UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA VEGETAL

Metagenomic mining of pathogenicity and antibiotic resistance traits across human populations worldwide

Pedro Miguel Agostinho Escudeiro

Mestrado em Bioinformática e Biologia Computacional
Especialização em Biologia Computacional

Dissertação orientada por:
Professora Doutora Teresa Nogueira
Professor Doutor Francisco Dionísio

2016

Acknowledgements

I would like to thank:

- ◇ **Teresa Nogueira, Ph.D.**, for accepting me into the research project that culminated in this present dissertation; for being my mentor; for introducing me to the exciting field of metagenomics and Unix Shell scripting; for teaching me so much about bioinformatics; for acquainting me with the inner workings of the academic world, and for the endless patience, compassion, support and encouragement bestowed upon me throughout this life-changing journey.
- ◇ **Francisco Dionísio, Ph.D.**, for providing me with the opportunity to enroll in this research project; for being exceptionally enthusiastic and supportive; for the kind words of confidence, reassurance, and inspiration; for his straightforwardness and determination every time I wavered; and for explaining to me what I didn't understand, and what I thought I did.
- ◇ **Joël Pothier, Ph.D.**, for steadily replying to my ceaseless stream of emails with unprecedented forbearance; for teaching me a lot about biostatistics; for being deeply interested and involved in this research project; for constantly putting my critical thinking to test; for welcoming me at the *Atelier de Bioinformatique*, and for his altruism and thoughtfulness.
- ◇ **The *Atelier de Bioinformatique* (ABI) team**, for receiving me in their *headquarters* with amazing hospitality; for allowing me to present this research project's findings amidst their peaceful and extremely astute environment; and for all the constructive criticism.
- ◇ **Eduardo Rocha, Ph.D.**, for acquainting us with one of the addressed databases, and for providing extremely valuable feedback throughout the present research project.
- ◇ **Octávio S. Paulo, Ph.D.**, for his promptness, and generosity, in granting me access to his group's server.
- ◇ **Francisco Pina-Martins, M.Sc.**, for answering all my questions regarding the latter server (and Unix in general) with the utmost kindness and patience.

Aos meus Pais

Resumo

O microbiota humano (a soma de todos os microrganismos que colonizam o corpo humano) é composto aproximadamente por 1e+14 células bacterianas, que abrangem vários taxa, e colonizam principalmente a pele, mucosas, tecido conjuntivo, e o tracto gastrointestinal, nomeadamente o cólon. O somatório de todos os genomas microbianos que lhe dizem respeito é frequentemente denominado microbioma.

O conjunto de genes que codificam virulência (eventualmente conferindo patogenicidade à bactéria) são frequentemente codificados em elementos genéticos móveis. Deste modo, muitos factores de virulência (VF) bacteriana conseguem ser facilmente disseminados em populações bacterianas por transferência horizontal de genes (movimento de material genético entre células), convertendo bactérias mutualistas ou comensais em potenciais patogenes.

Analogamente, a colecção de genes cujos determinantes (produtos génicos) conferem resistência a antibióticos (AR), existentes tanto em bactérias patogénicas como não-patogénicas, também se apresenta repetidamente codificada em elementos genéticos móveis, os quais sob pressão selectiva, se conseguem disseminar por entre comunidades bacterianas através do processo de transferência horizontal de genes, atravessando por vezes as barreiras taxonómicas de espécie e género. Esta característica permite que a comunidade sobreviva, e persista como um todo, comportando-se como um reservatório de genes conferentes de resistência.

Microbiomas ambientais, como os que se encontram presentes no solo, são descritos como reservatórios abundantes de genes de resistência a antibióticos. Estes codificam para determinantes de resistência a todas as classes de antibióticos descritas até hoje.

Apesar da antibioticoterapia ser direcionada a bactérias patogénicas, esta também afeta muitas espécies bacterianas não-patogénicas que fazem parte do microbiota dos indivíduos sujeitos a este tipo de terapia medicamentosa. Efeito que também se verifica em bactérias ambientais que se encontrem expostas a este tipo de pressão selectiva consequente de más práticas agrárias e da pecuária, ou simplesmente poluição antropológica.

Por conseguinte, o microbioma humano detém genes de resistência passiveis de transmissão a estirpes patogénicas, tornando-o num repertório de determinantes de resistência a antibióticos altamente diversificado.

O estilo de vida virulento, característico de bactérias patogénicas, tem sido consecutivamente associado a fenótipos de resistência a antibióticos. No entanto, esta associação nem sempre tem sido direta e previsível. Por um lado, o crescente uso de antibióticos tem vindo a seleccionar bactérias detentoras de fenótipos resistentes, sejam elas patogénicas ou não, criando reservatórios genéticos de resistência em diversos microbiomas. Porventura não é claro se a diversidade de genes conferentes de virulência se correlaciona com a diversidade dos que conferem resistência a antibióticos. Em contrapartida, existem inúmeros relatos bibliográficos de estirpes altamente virulentas e multi-resistentes (resistentes a mais do que uma classe de antibiótico) que têm vindo a disseminar-se por todo o globo.

Tendo em conta que atualmente existe uma grande disponibilidade de antibióticos, e em alguns casos, administração não supervisionada, podemos concluir que os microbiotas humanos, bem como os seus respectivos microbiomas, estão sujeitos a diferentes graus de pressões seletivas impostas pelos referidos compostos.

Neste contexto, podemos inferir que para alguns patogenes, em ordem a sobreviver e colonizar o hospedeiro, codificar apenas para virulência pode não ser suficiente, se se encontrarem na presença de antibióticos. Por outras palavras, sob o efeito de pressões seletivas impostas pela administração

de antibióticos, a seleção de elementos genéticos móveis que codifiquem para resistência aos referidos compostos juntamente com características genotípicas que confiram virulência irá ocorrer, tendo como consequência a sua disseminação dentro de comunidades bacterianas pertencentes ao microbiota humano.

Tanto quanto sabemos, não existem registos bibliográficos sobre a dinâmica evolutiva que dita a epidemiologia de genes conferentes de resistência a antibióticos, e os que conferem virulência, colectivamente. Assim sendo, a presente dissertação pergunta se sob o efeito de pressões seletivas exercidas por antibióticos, os determinantes de resistência e de virulência se encontram co-representados tanto em diversos microbiomas ambientais, como em microbiomas provenientes do trato gastrointestinal humano.

Deste modo, foram escolhidos metagenomas a fim de abordar esta temática por várias razões. A mais preeminente prende-se com o facto das bactérias serem organismos sociais, vivendo em comunidades. Um metagenoma corresponde à panóplia de material genético isolado de uma comunidade, e posteriormente sequenciado, pelo que caracteriza o repertório completo de genes envolvidos em processos metabólicos, fisiológicos e ecológicos, como por exemplo, na adaptação ao ambiente pelas comunidades microbianas presentes na respectiva amostra sequenciada.

Subsequentemente, a prospecção de genes em metagenomas surge como uma metodologia fidedigna no que toca ao estudo das pressões seletivas a que uma dada população bacteriana foi sujeita, assim como ao estudo da co-seleção de características genéticas do microbiota amostrado como um todo.

No presente trabalho utilizamos 64 metagenomas ambientais, referentes a 12 biomas diversos, bem como 110 metagenomas do trato gastrointestinal humano, originários de indivíduos pertencentes a várias comunidades dos Estados Unidos da América, Venezuela, e Malawi, caracterizando várias faixas etárias, diferentes culturas, hábitos alimentares, bem como diferentes graus de acesso a saneamento básico, a cuidados médicos e antibióticos. Todos os metagenomas encontram-se publicamente disponíveis para download no servidor MG-RAST, tendo sido descarregados em ficheiros individuais em formato FASTA, nos dias 3 de Abril de 2015 (metagenomas do trato gastrointestinal humano) e 17 de Novembro de 2015 (metagenomas ambientais). Cada ficheiro compreende sequências proteicas previamente agrupadas a 90% de homologia pela *pipeline* de formatação de ficheiros do servidor MG-RAST, contendo assim sequências traduzidas não redundantes, e representando deste modo a diversidade proteica de cada metagenoma.

O programa BLASTP foi utilizado a fim de inferir homologia de sequências proteicas envolvidas no fenómeno de resistência a antibióticos, bem como de virulência, de entre os metagenomas escolhidos, fazendo uso de duas bases-de-dados públicas: Resfams AR *Proteins database* (base-de-dados de proteínas bacterianas conferentes de resistência a antibióticos), e VFDB (base-de-dados de proteínas bacterianas envolvidas em virulência). Este programa permite inferir homologia entre sequências comparadas por via de um algoritmo de alinhamento de sequências derivado do algoritmo original de Smith-Waterman.

De entre os vários critérios de seriação aplicáveis foi escolhido um limiar de *E-value* = 1e-15, com um filtro posterior que apenas considera o melhor alinhamento para cada sequência proteica, e que satisfaça os requisitos mínimos de 60% de homologia sob 75% de alinhamento entre sequências comparadas. Ulteriormente ainda se removeram os alinhamentos resultantes de sequências proteicas que tanto eram homologas de determinantes de resistência a antibióticos como de factores de virulência, de modo a eliminar um viés na análise estatística consecutivamente implementada.

Seguidamente, de modo a aferir o tipo de relação entre os caracteres genéticos em causa, as contagens das diferentes sequências proteicas homologas de determinantes de resistência a antibióticos

(ARd) foram correlacionadas numa primeira fase com o número de sequências proteicas presentes em cada metagenoma, previamente agrupadas a 90% (tamanho do metagenoma), procedendo de igual forma para as contagens de diversidade de sequências proteicas homologas de factores de virulência (VFd), para ambos os grupos de metagenomas considerados. Posteriormente as contagens ARd e VFd foram correlacionadas entre si, após a estandardização das mesmas. Em ordem a medir o grau de associação entre as correlações previamente descritas recorreu-se a medidas estatísticas como o coeficiente de correlação e o *rho* de Spearman, bem como o seu *P-value*. Foram também geradas todas as possíveis associações entre as contagens de ARd e VFd para subfamílias proteicas funcionais caracterizadas nas bases-de-dados mencionadas, efetuando uma correção aos *P-values* resultantes do *rho* de Spearman pelo procedimento de Benjamini-Hochberg.

Em ordem a testar a dissemelhança entre médias provenientes dos rácios estandardizados de ARd/VFd em função da idade dos indivíduos pertencentes ao grupo de metagenomas do trato gastrointestinal humano, foram aplicados *Welch Two Sample t-tests* por pares, de acordo com os respectivos países de origem.

Os nossos resultados mostram que os determinantes de resistência a antibióticos, bem como os factores de virulência, se encontram amplamente disseminados tanto em microbiomas ambientais, como em microbiomas do trato gastrointestinal humano pertencentes aos 110 indivíduos saudáveis originários de países diferentes.

Em segundo lugar, também sugerem que, apesar das comunidades bacterianas ambientais possuírem maior variação de ARd e VFd tendo em conta o tamanho dos metagenomas, as comunidades que habitam o trato gastrointestinal humano detêm uma dependência linear muito forte no que toca à distribuição de ARd de acordo com o tamanho dos metagenomas, e uma relação linear forte entre VFd e o tamanho dos mesmos.

Adicionalmente constatamos que as contagens estandardizadas de ARd e VFd apresentam uma correlação muito forte entre si nos metagenomas de origem ambiental, sendo que estas contagens também se mostraram fortemente correlacionadas no grupo de metagenomas provenientes do trato gastrointestinal humano. Entre os metagenomas do grupo anterior, os referentes aos indivíduos originários dos Estados Unidos da América, apresentam uma ampla diversidade de associações, ao passo que as amostras provenientes de indivíduos Venezuelanos não possuem uma associação estatisticamente relevante. No entanto, os metagenomas pertencentes a indivíduos Malauianos retratam a correlação e associação linear mais forte de entre os três países amostrados, possuindo também duas vezes mais contagens de ARd por VFd que os outros dois países.

Referindo-nos ainda ao mesmo conjunto de metagenomas, conseguimos verificar que as contagens estandardizadas de ARd e VFd aparentam decrescer com o aumento da idade dos indivíduos, enquanto que os rácios ARd/VFd afiguram-se relativamente estáveis, evidenciando um incremento comum durante o primeiro ano de vida.

Relativamente a todas as associações geradas entre subfamílias de proteínas que codificam para AR e as que codificam VFs, aquelas que se relacionam com o envelope celular bacteriano apresentaram as melhores correlações e estatísticas correspondentes.

É de salientar que os resultados descritos neste trabalho apenas fornecem evidência para a co-representação de determinantes de AR e VF entre os metagenomas ambientais e do trato gastrointestinal humano que foram amostrados. Visto que a inclusão ou proximidade de determinantes AR e VF nos mesmos elementos genéticos móveis não se prende com os objectivos da presente dissertação, os nossos resultados não podem confirmar que a mobilização dos demais determinantes esteja a ocorrer conjuntamente. De qualquer modo, a natureza co-representativa dos nossos resultados

reforça a noção, bem como a hipótese de co-seleção dos referidos determinantes.

Palavras-Chave: Resistência a antibióticos; virulência; microbiomas humanos; microbiomas ambientais; metagenómica

Abstract

Genes contributing to the pathogenicity of a particular bacterial species are often grouped in pathogenicity islands, and encoded on mobile genetic elements such as plasmids or phages, as happens with some genes coding for resistance to antibiotics. Pathogenic bacteria have gradually become resistant to antibiotics as a result of intense selective pressure they are subjected to. Here we provide evidence that further reinforces the hypothesis on which, under antibiotics selective pressure, resistance and virulence traits are co-selected amongst bacterial communities naturally occurring in the human gut microbiome. Through means of metagenome mining, we have studied 64 environmental metagenomes from 12 diverse biomes, as well as 110 human gut metagenomes issuing from individuals belonging to different human populations across the world, having contrastive cultural, dietary and sanitary lifestyles, along with different medical access to antibiotics. Our results demonstrate that there is a great diversity of antibiotic resistance (AR) and virulence factor (VF) genetic traits amongst metagenomes in general. In the human gut there are less AR and VF genetic traits than in more versatile environments, yet the correlations between the latter determinants are still strong, advocating that in the human gut microbiome, there appears to be co-selection of these traits, remaining well established and long lasting in the foregoing host's microbiome. In the USA human gut metagenomes there are a few examples of AR determinants per VF accumulation, suggesting a possible consequence of antibiotic consumption abuse. In Malawi, a very poor African country, where there is a high prevalence of unattended antibiotic consumption, the correlation between AR determinants and VFs is very strong, as opposed to the scenario portrayed by the metagenomes pertaining to Amerindians (native populations of the Venezuelan Amazon) where there is neither reports of pharmaceutical-grade antibiotics consumption nor correlation at all, thus allowing us to link this effect to antibiotic exposure. Furthermore, the best correlations gathered between AR and VF protein sub-families amidst both metagenomic cohorts relate to the bacterial cell envelope, namely multidrug efflux pump components (AR determinants), along with secretion systems, adhesion proteins and iron-acquisition systems (VFs), most of which are known to be encoded within mobile genetic elements.

Keywords: Antibiotic resistance; virulence; human gut microbiome; environmental microbiomes; metagenomics

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1 Brief conceptualization on the human gut Microbiota

Microbial communities constitute the bedrock of life on Earth, and have seemingly been crucial for the very evolution of life, as we perceive it [1]. Fourteen years ago, Carl R. Woese conjured that the first steps of evolution did not arise from selective pressures imposed on intralineage variations nor on individual genealogical traces, but rather from imposing these pressures to the community as a whole, the ecosystem [1]. Quite possibly, these first microbial communities, although primitive in nature, bore striking resemblances to those of more primeval microbes found nowadays throughout the biosphere. It is likely that the extent of innovations required by microorganisms to become divergent in terms of cellular lineages, and populate the vastness of ecosystems present in today's biosphere, required whole communities to exchange molecular information. Henceforth, one of many milestones of microbial evolution lies in community dynamics, as a purported singular unit [1].

Fossil records date the occurrence of one of the first evidences of spatially organized microbial communities as soon as 3.25 billion years ago [2]. On the other hand, multicellular eukaryotic life is thought to have emerged around 1.2 billion years ago [3]. The two previous statements provide us with an insight into the extensive period of time on which both microbial communities and multicellular life forms interacted, and quite possibly wrought each other's evolutionary pathway [4]. Indisputable evidence of such associations can be ascertained since diverse multicellular eukaryotes, namely vertebrates, maintain evolutionary conserved responses to microbial colonization [5,6]. Amongst the previously stated are humans, whose genome is in due part product of the prime factors that behest the associations we have developed with our microbial neighbors, suggesting that these associations are not only due to mere cohabitation, but also due to the exchange of information (e.g.: changes in nutrient intake) between the host and the microbial communities inhabiting it [5].

The human body has been regarded to be inhabited by more bacteria than the number of mammalian cells it comprises by tenfold, and it has been further implied that such bacteria enclose 100-fold more unique genes than our very own genome [4,5] (where a genome is the collective sum of all the genes present within a given organism). The coinage of the term microbiota is commonly attributed to Joshua Lederberg, who defined it as "the ecological community of commensal, symbiotic, and pathogenic microorganisms that literally share our body space" [7]. Nevertheless, one can thusly apply this definition to the microbial communities present in other animals, as well as plants. Several other definitions of the human microbiota have risen since then, and the same concept can be formally conveyed as the totality of microorganisms that inhabit the specific ecological niches comprised by the human body, as a whole. Whereas the collective sum of their respective genomes is commonly named microbiome [8]. The concept of microbiome is usually applied throughout the literature in a much broader sense, owing to the fact that a microbiome can be of environmental nature [9]. Nonetheless, although these terms are sometimes used interchangeably, as if they were synonymous, semantically speaking, many authors make a clear distinction between the definitions of microbiota and microbiome, in order to describe either the communities of microorganisms themselves, or the collection of their genomes, respectively [4,5,8]. For clarification purposes, the present dissertation will acknowledge and make use of such distinction.

Further drawing on the concept of microbiome, one should also be acquainted with the term metagenome. A metagenome, similarly to a microbiome, can be defined as the collective sum of genes directly sequenced from a given environmental sample. Hence, a metagenome can portray the

phylogenetic and metabolic diversity present within a microbial community, factor which can lead to its analysis, and further study, somewhat similarly to that of a singular genome. This way of approaching microbial community analysis, named metagenomics, was first quoted by Jo Handelsman, Jon Clardy, Robert M. Goodman, Sean F. Brady, and other colleagues, who also coined the term "metagenome" in 1998 [10].

The vast majority of the foregoing prokaryotic cells that compose the human microbiota is encompassed within our gut (i.e.: the human gastrointestinal tract) [5], particularly the colon where concentrations approach 1e+11 ~ 1e+12 cells/ml, which in due term pinpoint it as the most densely populated microbial habitat recorded thus far [11]. Human gut's microbial community is noteworthy for its singularity in terms of phylogenetic diversity, given the inner workings of constant selective pressures exerted indirectly by the host (such as the physical and chemical setting; nutrient intake; host immune system; fewer biochemical niches than several environments, like the soil) [4] prune this diversity by members of just two divisions of bacteria – the Bacteroidetes and Firmicutes – and one member of Archaea, *Methanobrevibacter smithii* [12]. However, the human gut microbiota preferentially assorts itself into three well-defined bacterial community configurations (i.e.: clusters), with distinct metabolic and phylogenetic signatures. These cluster assortments have been coined by Arumugam and colleagues as enterotypes [13]. Moreover, differences concerning human gut microbiota diversity can arise from inter-individual and even intra-subject comparisons, the latter being due to surface-dwelling and lumen-dwelling microbial populations being contrastive and performing different tasks within the ecosystem (e.g.: stool versus mucosal communities) [12]. As such, these assortments may not exhibit an explicit arrangement throughout the major subdivisions of the colon and its lumen, but can instead reveal the presence of specialized microbiological niches [12]. The diversity reported insofar may very well be a consequence, at a lower phylogenetic level (e.g.: genus, species), of persistent colonization efforts exerted by the most bountiful representatives within the same biochemical niches [14].

Amongst the autochthonous microbiota's diversity underlies a complex network of interrelationships, between the host and the microbiota, and amidst the microbiota itself. The relationships that make up this intricate system are resultant of a prolonged and complex co-evolutionary process, as stated above [4]. These interconnections, between the host and its respective microbiota are suggested to be mutualistic rather than commensal, mainly because the term commensal – as duly noted by Bäckhed et al. [5] – typically conveys a noncommittal stance when referring to the undeniable beneficence of most individuals pertaining to this microbial community, to our own gain in terms of fitness, or to that of other community members [5]. From our point of view, the host, it comes as preeminent to fathom the beneficial labors our gut microbiota undertakes, which conclusively end up affecting our fitness as individuals. Although human beings can survive in the absence of gut microbiota [15], the latter is crucial to the host's innate digestive processes and nutrient acquisition. The gut microbiota generates several nutrients derived from substrates that would be otherwise indigestible by the host's gastrointestinal tract, namely xyloglucans and short-chain fatty acids [16]. It has also been summarized by LeBlanc et al. [17], that the gut microbiota's constituents are able to synthesize some vitamins, like vitamin K, and most of the B-complex vitamins, such as biotin, cobalamin, folates, nicotinic acid, panthotenic acid, pyridoxine, riboflavin and thiamine [17]. Focusing on some other, more conspicuous roles our gut microbiota plays, when it comes to our gain as host, we have: cardiovascular protection, host fat storage regulation, immunomodulation, inhibition of pathogens, normal gut motility, protection against injuries inflicted upon epithelial cells, and even stimulation of gut angiogenic processes [12,18]. From the gut microbiota's standpoint on this two-way relationship, one should mention that the microorganisms that encompass it are provided with (almost) constant nutrient

flow (particularly polysaccharides), protection from prospective competitors, and a specialized anoxic environment [5].

Emphasizing on the reasoning that has been built up thus far, we have that the human body (particularly the gut) and its microbiota have underwent millions of years of co-evolutionary processes [4], while developing a mutualistic partnership where both the microorganisms and the host are interconnected with one another in several ways, where the first is reliant on the latter on several aspects [5]. Firstly, the diverse microbial community in the gut is dependent on its host for an ecological environment in which to survive and is therefore also susceptible to the underlying environmental factors exposed to the host. Secondly, the microorganisms within the human gut are also proven to be more adaptable and to change in a faster manner than its host genome, potentially providing it with quick adaptive advantages [5]. The preceding remark portends that each member in this mutualistic relationship may find itself under different selective pressures: the microbiota is susceptible to changes when its host diet is reformed, when environments are briskly altered, or when bestowed upon with chemical agents, such as antibiotics. On the contrary, the human genome is subject to adaptation only either through mutational or epigenetic events underwent between generations [19].

However, the communion between host and its microbiota is not always amicable. Despite the fact that the aforementioned relationships are in its overwhelming majority mutualistic in nature, there are instances where pathogens arise, disrupting and/or taking advantage of the community dynamics while usually causing a disease state to the host [4,20]. Still, from an enteric outlook, there seems to be a distinction between the classical view of pathogenesis where a recent (i.e.: non-preexistent), acute disease state, is caused by the arrival of an alien pathogen that has infiltrated the intrinsic microbiota, versus dysbiosis – also referred to as dysbacteriosis. The former designation can be fairly described as a disruption in the equilibrium that lies between the putative species of beneficial and potentially opportunistic harmful bacteria [21]. Sometimes this disequilibrium is chronically perpetrated so that the whole gut microbiota can become harmful to its host, as happens with several maladies such as irritable bowel syndrome, inflammatory bowel disease, nonalcoholic fatty liver disease and obesity [22]. Yet, the question of whether dysbiosis is a cause or a consequence of some of the mentioned illnesses remains elusive. Elaborating on this notion of dysbiotic pathogens, one can view them as pertaining to an assemblage in which the whole microbial community is regarded as "harmful". In other words, it is not as beneficial to its host as a "healthy" community would be. In this community, no sole microbial cell can be witnessed as a pathogen by itself. Instead, the community as a whole composes a circumstantial hazard that adds up to a disease state. Additionally, this microbial community should be acknowledged as pathogenic always within the ambiance of other emergent factors, such as host genome, diet, and risk behaviors [4]. This dissertation will acknowledge the distinctions made heretofore and will refer to pathogenicity in a "classical" sense, as opposed to dysbiosis, unless stated otherwise.

1.2 Bacterial Pathogens and Virulence Factors

A pathogen is defined as a microbe that can cause damage to its host. And once again, this definition can encircle pathogens as parasites in a classical sense, as well as opportunistic ones (i.e.: dysbiosis). The damage sustained by the host, usually manifested as a disease state, may either be consequent from actions enforced by the pathogen itself or the host's immune system. Subsequently, virulence can be deemed as the relative capacity that a pathogen possesses in order to cause damage to a host, minding that virulence factors (VFs) – also known as virulence determinants – are the components of a pathogen that enable it to cause damage upon the host [20].

Pathogenic bacteria have always been a notorious threat to public health and Humankind's survival, causing misery and death in its wake throughout recorded history [23]. Perhaps the most blatant example of such statement was the Black Death pandemic (also known as Black Plague or Bubonic Plague) that laid waste to Europe in the 14th century, resulting in the death of tens of millions of people [24]. This plague continued to plunder the Continent far into the 18th century, with countless epidemic resurgences [24]. Several clones of the *Yersinia pestis* bacterium have been held accountable for said devastation after DNA analysis of human skeletons dating from the 14th-17th centuries, pertaining to presumable victims from northern, central, and southern Europe was first made public in 2010 [24]. Nowadays there are still plenty of disease-causing bacteria, which often lead to epidemics that threaten society, and health-care systems worldwide. One of the most noteworthy bacterial pathogens that has been resurfacing countless times in epidemic waves, either in developed or developing countries, is *Vibrio cholerae*, which causes Cholera, with the remarkable peculiarity of only affecting human beings [25]. Since 1817, seven documented cholera pandemics have tormented mankind, although the disease itself has been depicted as early as the 5th century BC [26]. Despite the fact that the state of the disease can be often classified as moderate or asymptomatic, some of the afflicted patients show profuse watery diarrhea and vomiting. Severe disease-states display tremendous loss of bodily fluids, quickly resulting in dehydration, hypovolemic shock (drastic volume decrease of blood plasma, resultant from acute dehydration), and subsequent death [25]. The predominant symptoms the disease portrays are a blunt consequence of two major VFs expressed by *V. cholerae*: the cholera toxin (CTX), and the toxin co-regulated pilus, the last being indispensable for the pathogen's further colonization of the intestine [25].

Outlining the fact that this bacterial pathogen is human-exclusive, the previous statement poses as a quintessential illustration that human-bacterial co-evolution not only happens at the host-microbiota stage (like it was discussed in the previous subsection), but also at the host-pathogen level [27]. This host-pathogen relationship can be perceived as an everlasting biological arms race between microbial pathogens and humans, on which, rounds after rounds, pathogens develop new attack strategies, in the shape of VFs, and hosts counter defend these offensive strikes, through means of immunological innovations. This drives forward co-evolution on both parts, and ensues biological novelties [28]. Construing on the highlighted relationship, it comes as logical to ascribe infection as a property of the interaction between the host and the bacterial pathogen, being that non-achievement of infection is as much of a host feature as it is also an attribute of the foregoing pathogen [29]. Recollecting the host specificity that *V. cholerae* displays, it's only rational to ground that failure to infect species other than its host (i.e.: other than humans) might well be the outcome of a lengthy evolutionary process that lead to the specialization of this particular bacterial pathogen on its very own source host species [29]. Like *V. cholerae*, many other bacterial pathogens are exquisitely suited to exploit their designated host. Their ecological niche becomes predefined by the biochemical milieu the host has to offer, in this case the nutrient rich environment of the human gut [30]. As so, bacterial pathogens are selected on the basis of exhibiting environmental responses on their side, as well as molecular traits that confer adaptation, which in due course, if successful, allow them to persist within the host [30]. One of the main selective pressures that continuously shape the phenotypic behavior of bacterial pathogens is the host immune system. Host immunological responses, following contact with a bacterial pathogen, are themselves adaptive in nature, and as such, they seek to neutralize or restrict bacterial replication (i.e.: proliferation). Therefore, bacterial pathogens deemed successful must either avoid or adapt to the host's ever-evolving immune defenses. Anyhow, in order to persevere within a niche over extensive periods of time, the pathogen must be able to survive within the individual host and also be apt to infect other hosts. To

achieve ecologic success (i.e.: persistence in its niche within the host population), bacterial pathogens require mechanisms that provide them with survival skills within their designated hosts, and transmission between the latter [30]. These mechanisms are, once again, designated as virulence factors (VFs).

Bacterial pathogens are in a general sense, a true scourge on humanity. Yet they fall nothing short of remarkable from the bacteriological viewpoint, as they can be termed professional in their capability to display virulent behavior, withstanding the presence, and actions, of either innate or acquired host defenses, and still be able to cause damage as well as induce a disease-state on their host [31]. In 1996, Stanley Falkow pointed out that the main aptitudes required of a pathogenic bacterium in order to be successful in its endeavors are (i) to enter the host body; (ii) to secure an exclusive niche; (iii) to evade or undermine the host's immune system (innate or acquired) responses, as well as overthrow other competing microbes; (iv) to attain indispensable nutrients to its survival and/or proliferation; (v) to reproduce and/or persevere; (vi) to cause damage and/or to induce a disease-state to its host; and finally, (vii) to exit the present host while still being able to cause infection to new susceptible hosts [32]. On a side note, despite the fact that the summary authorship of these bacterial pathogens' core competencies is usually attributed to Falkow, it was Smith who many years ago, conceived and advanced a first synopsis of these basic concepts, still valid as of today [33].

Notwithstanding this brief curiosity, one can further realize that virulence genes – which after being translated into protein are deemed virulence factors (VFs) or virulence determinants – genetically encode the core competencies needed of bacterial pathogens. These determinants are molecules (e.g.: proteins) produced by the former microbes with the intent of overturning core biological functions of the host. Along these lines, the pathogen aims at modifying basic functions of the host, as to uphold its survival and/or propagation. This overturn is frequently a product of the characteristic interaction between the pathogen's VFs and proteins encoded by the host, as well as various other molecules present within the host's system [31]. Subversion of the host's homeostasis, if successful, results in infection and sometimes a concomitant disease state, in which both the host and the infection-causing pathogen engage in a battle for survival. Just as the host possesses refined immunological responses against infection – evolutionarily conserved throughout millions of years [34] – the infection-causing pathogens also bear correspondingly intricate molecular mechanisms to offset, and further along, overturn the host immune responses set to neutralize them [35]. This vast arsenal of molecular machineries assists them in the achievement of their ulterior goals (e.g.: subsistence; propagation), bringing to mind that the ownership of such mechanisms (i.e.: VFs), encoded within the pathogen's genome, is the determining factor that allows us to tell apart virulent microbes from non-virulent ones [31]. Hence, the degree of pathogenicity of a certain pathogenic species can be attained by its relative abundance of VFs encoded within the pathogen itself, as well as their function regarding the host-pathogen relationship.

The contribution of a specific gene to a bacterial pathogen's virulence can be established in the light of the molecular Koch's postulates [36]. Falkow established these postulates in 1988, which were based on the original Koch's postulates. The original Koch's postulates were formulated as to pinpoint the causative relationship between a probable pathogen and a specific disease. For purposes of clarification, the molecular Koch's postulates, as originally established by Falkow are quoted as follows: (i) "The phenotype or property under investigation should be associated with pathogenic members of a genus or pathogenic strains of a species"; (ii) "Specific inactivation of the gene(s) associated with the suspected virulence trait should lead to a measurable loss in pathogenicity or virulence"; and (iii) "Reversion or allelic replacement of the mutated gene should lead to restoration of pathogenicity" [36]. Furthermore, a given gene is not required to fulfill all three postulates, but its relative contribution to the virulence of a particular bacterial pathogen can be acknowledged as bearing more weight if it is present

in the latter but absent from closely related non-virulent bacteria (e.g.: different strain); if rendering the given gene inactive (e.g.: by means of mutation) results in loss of the corresponding virulence mechanism; and if by replacing the inactivated gene with an exact copy of the original one results in rehabilitation of the virulent phenotype belonging to the preceding bacterium [37]. Virulence genes are frequently found on transmissible genetic elements such as bacteriophages (viruses that infect and replicate exclusively inside bacteria), plasmids (small, circular, double-stranded DNA molecules that are independent from the chromosomal DNA), and transposons (DNA sequences that can change their position within a genome)[32]. In addition to the previous, virulence genes are also found within the bacterial chromosome, were they often appear clustered together in contiguous domains that go by the name of pathogenicity islands [38]. These clusters usually enclose sets of specific genes, the translated products of which contribute to a respectively specific virulence function, just as the main aptitudes required of pathogenic bacteria as summarized by Falkow [32]. Bacteria, unlike multicellular eukaryotic organisms, can exchange genetic information between dissimilar taxa in diverse environments through means of a process termed horizontal gene transfer (this process shall be discussed in greater detail in the following subsection) [31,39]. Horizontal gene transfer (HGT) has been implicated in a swift and broad dissemination of virulence mechanisms amongst diverse pathogens, such as those encoded within pathogenicity islands that encircle analogous genes and perform similar functions [31]. Another important aspect that concerns virulence genes is that they do not exhibit constitutive expression in a regular fashion, but are alternatively only expressed after establishing contact with the host, or after invasion of the latter [40]. Moreover, the expression of these genes in the host (i.e.: *in vivo*) is mainly dependent on the pathogen's capacity to perceive its surrounding environment and recognize that it has established contact with the host [31].

A predominant topic in the abiding discussion of bacterial pathogenesis is that most virulence conferring mechanisms act by subverting host biological processes, and its consequent homeostasis [41]. And, as previously explained, these virulence mechanisms, conferred upon bacterial pathogens as a result of the expression of virulence determinants, can be succinctly classified according to the core competencies they grant to the bacterial pathogen in question, like those summarized by Falkow [31,32]. Despite the fact that there are various ways of classifying virulence mechanisms and VFs, the present dissertation will make use of a broad classification focusing on the main types of both. As such, VFs can be broadly classified by five different main virulence mechanisms: (i) adhesion; (ii) invasion; (iii) secretion systems; (iv) toxins; and (v) nutrient acquisition (of which only iron acquisition systems shall be disclosed) [42]. However, despite the fact that the present dissertation shall resume all five main virulence mechanisms, only adhesion, secretion systems, and iron acquisition will be examined in greater detail.

It is well established that an essential stage in any given bacterial infection is the capacity to adhere to the surfaces in contact with, or enclosed within the host, bearing in mind that a prospective pathogenic bacterium can also adhere to abiotic materials, such as those essential for life support of critically ill patients (e.g.: endotracheal tubes). This peculiar characteristic has been thought to be involved in the dissemination of nosocomial infections [43]. Therefore, most bacterial pathogens need the ability to attain intimate contact with host surfaces (e.g.: extracellular matrix) in order to achieve prosperous colonization, and wherefore induce a disease-state. The VFs responsible for carrying out such tasks are commonly named adhesins [42]. Adhesins can fall into two particular functional categories: initial contact and/or colonization of the surfaces enclosed within the host through receptor-specific interactions with host cell receptors; and close bonding of the bacteria with host cell surfaces – event which ultimately leads to following invasion [31]. Bacterial adhesins show refined affinity for target molecules produced

by the host such as enzymes and immunoglobulins (key proteins of the host acquired immune system) [44]. This extraordinary level of precision is species and tissue specific, depending on the bacterial adhesin that is brought into play. The latter can be witnessed in several strains of enteropathogenic *Escherichia coli*, which express adhesins that only adhere to the intestinal epithelium of humans and pigs [45]. Minding the fact that a sole bacterial pathogen may express a vast array of virulence-associated adherence mechanisms, and that VFs implicated in adhesion to the host may play other roles in pathogenicity as well, the part a specific adhesin plays in the infection process of the host can be strikingly difficult to ascertain [42]. Regardless of the previous hindrance, two principal adhesin groups shall be approached as archetypes of typical host-pathogen adhesion mechanisms, the first being the type IV pili (T4P) multi-functional adhesins. Structurally speaking, type IV pili adhesins can be construed as polymeric molecular complexes, composed of thousands upon thousands of pilin protein subunits, which together compose utterly thin filaments, named pilus, that stand a few microns in length [46]. Type IV pili have been credited with several virulence related traits, including surface migration, biofilm (groups of bacteria stuck together being usually attached to surfaces) production, adhesion, avoidance of host immunological responses, signaling between individual cells, DNA transformation and the attachment of bacteriophages [46]. Concerning type IV pili adhesion properties to the host-cell epithelium, one can expose as an example that the expression of the type IV pili complex is required in the commencing stages of infection undertaken by the pathogenic species of *Neisseria spp.* (a wide genus of bacteria that thrive in the mucosal surfaces of several animals) in order for it to be capable of attachment to human epithelial cells [46]. The pilus receptor has been conjured to be the glycoprotein CD46, which spans the cellular membrane of all known human cells except erythrocytes (red blood cells). This last glycoprotein is known to take part in the activation of the immune system's complement (an enzymatic cascade that recruits antibodies and phagocytic cells) [47]. The second group of adhesins to be considered here as an important virulence conferring mechanism of adhesion are fibronectin binding proteins (FnBPs). Fibronectin (Fn) presents itself as a 440-kDa glycoprotein that settles amidst the extracellular matrix and bodily secretions of animals. It was the first extracellular matrix protein confirmed to act as a substrate for the adhesion of eukaryotic cells namely through the membrane-spanning receptor proteins called integrins [48,49]. On the pathogen's side, fibronectin binding proteins comprise a subclass of bacterial surface adhesins that bind to the host protein fibronectin [48]. Acknowledging the fact that these proteins comprise a whole subclass of adhesins, one can fathom that their structure varies depending on the bacterial pathogen species being taken under consideration. Since most of the knowledge pertaining to bacterial FnBPs has resulted from the study of proteins belonging to the Gram-positive bacterial pathogens *Staphylococcus aureus* and *Streptococcus pyogenes*, the first will be taken as an example [50]. In the former bacterium, one of the two most studied fibronectin binding proteins (FnBPA) is structurally comprised by two binding domains identified so far [51]. It has been established that the characteristic interaction between *S. aureus* FnBPs and the host's fibronectin, which lies within the extracellular matrix, is able to expedite the binding of the bacterium to host cell surfaces by taking advantage of fibronectin proteins previously bound to the host cell integrin $\alpha5\beta1$ [52]. The binding of *S. aureus* FnBPA to human integrin $\alpha5\beta1$ by means of the previously explained process has also been shown to facilitate bacterial invasion of host cells [52]. A more recent study has further determined that biofilm formation in a *S. aureus* methicillin resistant strain is essentially reliant on the activity of FnBPs, thus reinforcing the multipurpose abilities inherent to this kind of VF [53].

Healthy host epithelium poses as a highly efficient barrier to bacterial pathogenic invasion [31]. As such, an ability to infiltrate intact epithelial surfaces proves itself to be an essential trait for many specialized pathogenic bacteria, as it has been thought that invasive measures undertaken by bacterial

pathogens end up providing a sheltered cellular milieu for the bacteria to replicate or persist upon [42,54]. Contrastively, in pathogenic bacteria that are not that specialized in terms of invasiveness, their only means of doing so depends on wounds or defects in the host epithelium (e.g.: peritonitis resultant of prior perforation of the intestinal tract) [31]. Equivalently to the connotation bestowed upon VFs that confer adhesion mechanisms to the bacterial pathogen in question (i.e.: adhesins), the corresponding VFs that render a specific bacterium able of invasive properties are termed invasins. Anyhow, pathogenic bacteria capable of crossing intact epithelial surfaces mainly do so by means of breaching through cells (i.e.: transcellularly), rather than between them (i.e.: intercellularly). Transcellular invasiveness is either initiated by the host epithelium cells or due to pathogen resourcefulness. Pathogen initiated invasion happens through subversion of host innate cellular mechanisms, leading to the consequent internalization of the former [31]. One such example is the internalization of the enteropathogenic bacteria *Shigella flexneri*. This bacterium produces and later secretes an invasin, named IpaB, which in due course disrupts the phagosome (a vacuole that contains a particle enclosed within part of the cell membrane) and allows the bacterium to break away freely into the cytoplasmic space [55]. The same bacterium also produces a protein (IcsA) located at its rear pole, that's responsible for initiating actin (a globular microfilament-forming protein, present in almost every eukaryotic cell) polymerization, thus enabling *S. flexneri* to move throughout the host cell cytoplasm, following penetration of neighboring cells, aiding in this way the circumvention of host immune responses, and further tissue invasion [56].

Bacterial pathogen's secretion systems are without a doubt an extensively studied subject in the field of bacterial pathogenesis, being that the vast majority of bacterial VFs are either located on the bacterial cell's surface, or secreted by the former systems [42]. Secreted VFs can portray many roles in the promotion of bacterial pathogenicity. These roles vary between the enhancement of adherence to host cells, to the scavenging of nutrients present within a niche, to direct intoxication of host cells and further disruption of their native functions. A vast amount of bacterial pathogens use specialized protein secretion systems as to secrete VFs directly from their cytoplasmic milieu into host cells or the host environment [57]. Some of the systems pertinent to secretion are outstandingly homologous, not to mention that various other VFs that strike as seemingly unrelated often share common transport mechanisms. The trend seen up to this moment in naming secretion systems as virulence mechanisms, or more generally, as indispensable mechanisms to any bacterial lifestyle, has led to their classification into functional groups according to the transport pathways employed [42]. In this fashion, and according to a recently published review from the authorship of Erin R. Green and Joan Mecsas [57], the secretion systems functional groups established so far range from type I (T1SS), to type VII (T7SS), also comprising: Sec, Tat, SecA2, Sortase, and the Injectosome [57]. However, it has also been proposed that yet another system, the extracellular nucleation-precipitation (ENP) pathway, to be renamed as the type VIII secretion system (T8SS) [58]. Likewise, some of the Gld and Spr proteins originally titled as the Por protein secretion system (PorSS) – characteristic from the phyllum Bacteroidetes – have been recently titled as the type IX secretion system (T9SS) [59]. Nevertheless, not all of the secretion systems considered thus far are common to both Gram-negative and Gram-positive bacteria. Perhaps the two most relevant secretion systems, taking this introduction's purpose into account, are the type III secretion system (T3SS), and the type VI secretion system (T6SS), being aware that they're only found in Gram-negative pathogens. The T3SS is commonly depicted throughout the literature as a "needle and syringe"-like mechanism, due to its structure and also due to the fact that it allows Gram-negative bacterial pathogens to secrete a myriad of substrates across both the inner and the outer bacterial cell membranes, and sometimes directly into host's cells, fact which has led it to being referred to as injectisome [57,60]. The latter shouldn't be confused with the Gram-positive's Injectosome,

proposed to possess an analogous function, yet an unrelated structure, to that of the T3SS and T4SS of Gram-negative pathogens [57]. Nonetheless, the T3SS, or moreover the injectisome has been depicted as an analogous mechanism to that of the bacterial flagellum [60,61]. Even though both mechanisms are structurally and functionally different, they both comprise conserved machinery for protein transport, which led some authors to classify the bacterial flagellum as a T3SS mechanism [60,61]. In opposition to translocation-associated type III secretion systems (i.e.: injectisomes), the flagellar T3SS secretes almost exclusively components of the flagellum to the extracellular environment. Nevertheless, there are reports of virulent factors being secreted by this type of apparatus [62,63]. Concerning the structure of the T3SS from the viewpoint of a translocation-associated mechanism, it comprises three core units: a base complex, the needle component, and the translocon [57]. In animal pathogens, the T3SS needle component has an inner hollow core through which enables unfolded proteins to be transported [57,61]. Following contact with a host cell, believed to be sensed through the needle, the translocon and tip proteins create a channel through which these proteinaceous substrates are translocated. Regardless of the assumptions on how this mechanism works, more experimental evidence is needed to ascertain the mechanism by which translocation occurs [57]. Nearly all Gram-negative bacterial pathogens that comprise type III secretion systems mainly use them to transport effector proteins (VFs involved in the subversion of normal host cell functions) across a target host cell membrane [57,60]. As such, this secretion system plays an elaborate part in the pathogenesis of many bacterial genera known to be virulent, such as *Yersinia*, *Salmonella*, and *Shigella* [61], where the injection of protein effectors allows such pathogens to supersede host native cellular processes, enabling the preceding bacteria to settle an infectious niche, being it within the host cells or amongst host tissues [57,60]. Moving on to our next topic of discussion, we find ourselves with the T6SS, one of the last discovered secretion systems [57]. This secretion system has been reported to be reasonably well conserved amongst a broad number of Gram-negative species of bacteria, where its main function is to translocate proteins straight into targets, namely host cells and other competing bacteria, resembling a "firing" motion akin to the mechanism behind contractile bacteriophage tails [57,64]. Acquainting with the fact that the T6SS was only identified as of 2006, there are still many underlying factors regarding its structure and putative functions that are unknown as of today [57]. It was firstly identified as a new secretion system directly involved with pathogenicity in *Pseudomonas aeruginosa* and *V. cholerae* [64], being only later proposed that it also served other functions such as the mediation of interbacterial interactions, either intra- or inter-specific [64,65]. Type VI secretion systems are very large complexes, that can encode as far as 21 different proteins within a single contiguous gene cluster, thirteen of which seem to be conserved in all type VI secretion systems, where they are believed to play a structural part in the secretion machinery [57,65]. Structurally speaking, the T6SS is believed to be made of two main complexes in association with additional cytoplasmic elements: an assembly present in the membrane, including two proteins homologous to bacterial type IV secretion system determinants, and an assembly whose components bear a structural resemblance to the sheath, tube and tail spike proteins present in bacteriophages [65]. These two assemblies cooperate by an unknown mechanism in which, the contraction of the bacteriophage-like sheath structure drives an inner tube terminated by a membrane-puncturing spike against a target cell [64]. Thus translocating effector proteins across the envelope of the bacterial cell in question, and then farther through the outer membrane of a target cell [65]. Although T6SS is believed to be directly implicated in bacterial virulence, some authors propose that the role T6SS plays in the latter might quite simply be that of enabling bacterial pathogens to compete more efficiently with the host microbiota [65]. All of the above stated in regards to type VI secretion systems, namely its widespread presence and mediation of interbacterial interactions, has fueled a deep interest in this peculiar secretion system, and

its fascinating mode of action [64].

Many pathogenic bacteria secrete VFs in the form of toxins. These are potent substances, usually under the form of enzymes, which are more than enough to dictate the outcome of an infective process. The major symptoms of a disease caused by a particular toxin-producing pathogen can be witnessed by the sole injection of modest doses of the purified toxin being considered, substantiating their relevance in bacterial pathogenicity [42,66,67]. These VFs may be directly or indirectly toxic to host cells, phenomenon that has led to their classification into functional groups according to their respective mode of action [66]. It has been known for some time that a specific functional group of toxins, the exotoxins, are for the most part secreted bacterial enzymes, that kill host cells at deftly low concentrations [42]. Aiming our attention at exotoxins, it has been reported that the foregoing toxins can be administered to host cells via the routes by which secretion systems operate, like the previously described T3SS [68]. Exotoxins can target distinctive host cell types, however some explicitly target macrophages and neutrophils (key cells of the human innate immune system) directly undermining host innate immune responses. Providing in such a way a convenient environment for active proliferation and inferable perseverance of the pathogen [67]. This is the case with two toxins secreted by *Bordetella pertussis* (the pathogen behind whooping cough), pertussis toxin (PT) and adenylate cyclase toxin (ACT). Being observed in murine models that this "dynamic duo" displays completing purposes in the pathogen's virulence, attacking host immune cells distinctively [69]. PT is thought to act mainly whilst infection is still beginning to settle, by recruitment inhibition of host innate immune cells. On the other hand, ACT is thought to attack macrophages and neutrophils at a later stage, thwarting phagocytosis and the posterior destruction of bacteria as an end result [69].

As for the last virulence mechanism to be discussed, we find ourselves with undoubtedly one of the most fascinating virulence conferring systems reviewed so far: bacterial iron acquisition systems. The human body poses as a bountiful reservoir of essential nutrients, one of them being iron (Fe), taking the fourth place as one of the most abundant elements in our planet's crust. It's also the most plentiful transition metal present in the human body [70], minding that in bacteria, it plays an imperative role in a vast assortment of physiological processes, such as being the cofactor of many enzymes, being involved in DNA replication and transcription, as well as central metabolism in general [71]. Likewise, it's only rational to ascertain that countless bacterial pathogens have co-evolved with their host as to exploit this valuable resource [70], which although being the most represented transitional metal within the latter it's still classified as a micronutrient, being available in very low concentrations. As such, most bacterial pathogens use these low iron concentrations as a hint to trigger the activation of certain VFs [42], especially the ones that shall be addressed as follows. According to a very recent review authored by Jessica R. Sheldon and colleagues [72], the mechanisms by which pathogenic bacteria acquire iron within their mammalian host include: (i) the extraction and/or capture of heme (protein cofactor consisting of a ferrous – Fe^{2+} – cation incorporated within the center of a heterocyclic organic ring called porphyrin) associated iron from hemoproteins present in the host through the usage of either proteins secreted by the bacterium or receptors present in the cell's surface; (ii) uptake from iron-binding blood plasma glycoproteins (e.g.: transferrin and lactoferrin) through binding proteins specific of the bacterium's cell surface or by means of siderophore (chelating compounds with high-affinity for iron) secretion; and (iii) the obtainment of free inorganic iron promoted by ferric (Fe^{3+}) iron bacterial intake proteins (i.e.: reductases and associated permeases) [72]. From all the formerly pointed iron acquisition mechanisms, the ones that comprise heme as a target are particularly preferred by pathogens, mainly because heme amounts to roughly 75% of the total iron available within the host [72]. However, heme is consistently complexed with hemoglobin inside host erythrocytes. Thus, pathogenic bacteria

have evolved ways to capture heme from within red blood cells, for instance, the active secretion of hemolysins: exotoxins that cause the disruption, or lysis, of erythrocytes, via destruction of their cell membrane [73]. After the disrupted erythrocytes have released their hemoglobin to the extracellular milieu, it stands as finally available for further capture by bacterial pathogens able to express specific heme/hemoglobin-binding proteins [72]. A few bacterial pathogens, acquire extracellularly available heme by synthesizing, and later secreting, soluble heme-binding proteins known as hemophores, being the latter able to appropriate free heme as well as extracting it from host expressed hemoproteins [72,74]. This salvaged heme is transferred later on to specific heme-binding receptor proteins. These heme-binding receptor proteins are either localized on the outer membrane, or on the bacterial cell wall, in case the pathogen is Gram-negative or Gram-positive, respectively [72]. Two major types of hemophores have been reported until now, one pertaining to Gram-negative pathogens (HasA-type hemophores), and the other to Gram-positive ones (near iron transporter – NEAT – domain-containing hemophores) [72]. Another widely used mechanism of iron acquisition is the secretion of siderophores by bacteria. Albeit siderophore production is a well-established virulence mechanism, it can be found as an iron-scavenging tactic throughout the prokarya domain [42]. Despite this generalization, one should regard siderophore use by pathogenic bacteria as a relevant virulence-conferring mechanism. Siderophores are small iron chelating molecules secreted by bacteria as to bind Fe^{3+} with a higher affinity than that of iron-binding blood plasma glycoproteins [75]. Even though siderophores are usually small and characterized as possessing low-molecular mass, they're still too large to pass through non-selective porins (beta-barrel proteins spanning a cellular membrane that act as a pore for the diffusion of molecules) of Gram-negative bacteria. Taking this case into account, the transport of siderophores comes as energy-dependent, mediated by the activity of specific porin receptors, especially those that are TonB-dependent (a family of beta-barrel proteins) [70]. Gram-negative bacteria lack ATP and ionic gradients in their periplasmatic surroundings, rendering them unable to drive the transport of molecules across the outer membrane while making use of these gradients. Because of this deterrence, they rather rely on energy originating at the inner membrane from the proton motive force. This energy is later harnessed by a specific protein mechanistic complex (the TonB-ExbB-ExbD system), in order to mediate active transport across the outer membrane. Meanwhile, in the periplasm, substrate binding proteins (SBPs), belonging to the ATP-binding cassette (ABC) transporter family – in this particular system – identify the siderophore-associated Fe^{3+} complex, and conclusively commute it to the respective ABC transporter. SBPs are expressed by Gram-positive bacteria as well, although being membrane-bound. Just as the siderophores reach the bacterial cytoplasm, the siderophore-associated Fe^{3+} is freed from the complex by means of chemical reduction to Fe^{2+}, or through the siderophore's subsequent enzymatic degradation, culminating in the availability of unbounded Fe^{2+} to be used as a nutrient by the pathogenic bacterium [70].

1.3 Antibiotic Resistance

Human beings have been battling pathogenic bacteria with the aid of antibiotics for nearly 70 years, ever since the accidental discovery of the very first true antibiotic – penicillin – by Alexander Fleming in 1928, and it's ensuing mass production in 1942 [76]. Although some antibiotics had been discovered as of 1928, penicillin was deemed the first true antibiotic [76], since it possessed bactericidal (killed bacteria) properties, as opposed to bacteriostatic (inhibit the growth of bacteria) ones. As thoroughly discussed throughout the literature, antibiotics came into existence as a wondrous remedy, being acknowledged as one of the most successful drugs ever developed [77-79]. Besides

allowing the treatment of infections, the broad use of antibiotics made the implementation of novel clinical practices possible. Procedures like induced immunosuppression following transplantation or anticancer chemotherapy, massive surgery, and even catheterization of patients in intensive care units, are now feasible and commonplace. These patients are more susceptible to infections, meaning that these procedures can be safely implemented as long as infections are able be treated or prevented [79]. All antibiotics, or antibacterial drugs to be more precise, discovered thus far, can be succinctly classified according to their action's mechanism towards bacterial growth or physiological processes in general, and also to whether they exhibit bacteriostatic or bactericidal abilities accordingly. Being faithful to such classification scheme – and for clarification purposes only – we have antibiotics that induce (i) inhibition of cell wall synthesis – β-lactams (e.g.: penicillin), daptomycin and glycopeptides (bacteriocidal); (ii) inhibition of DNA synthesis – fluoroquinolones, metronidazole (bacteriocidal); (iii) inhibition of protein synthesis – macrolides, lincosamides, streptogramins, chloramphenicol, ketolides, oxazolidinones, tetracyclines and aminoglycosides (only the latter are bacteriocidal); (iv) cell membrane binding – polymyxins and lipopeptides (bacteriocidal); (v) inhibition RNA synthesis – rifamycins (mainly bacteriostatic); (vi) and those that inhibit folate (a member of the vitamin B complex) synthesis – trimethoprim and sulfonamides (bacteriostatic) [80,81].

Unfortunately, it soon became quite clear, right after the discovery of each of the formerly introduced classes of antibiotics, that some bacteria were developing resistance to them. Even though antibacterial drugs consistent rate of discovery had been mitigating this drawback for quite some time, during the past 50 years the pace at which they have been discovered has slowed down substantially, leaving us with an ever-increasing legion of antibiotic-resistant bacterial pathogens [77,78]. As such, acquisition of resistance by bacterial pathogens has compromised not only the treatment of the respective infections they cause, but also the safe implementation of numerous clinical practices that we have been taking for granted [79]. In light of microbial ecology and evolution, antibiotic resistance (AR) in bacterial pathogens can be defined as an adaptive trait, acquired after imposing selective pressures, consequential with prior introduction of therapeutic antibiotics into the pathogen's environment [78]. It goes without saying that not every antibiotic resistant bacterium is pathogenic; just as not every pathogen is antibiotic resistant. However, the aim of this present subsection is to focus on antibiotic resistant bacterial pathogens.

Drug-resistant strains first appeared in nosocomial environments – the very source of antibiotic administration. And although there are reports of sulfonamide-resistant *S. pyogenes* emerging in military hospitals as soon as the 1930s, it wasn't until the 1940s that penicillin-resistant *S. aureus* defied civilian hospitals in London, shortly after the introduction of this drug [80]. Nowadays, the greatest threat mankind faces concerning AR, is the one posed by multidrug resistant (MDR) bacterial pathogens, so-called superbugs [77,80]. In 2007 Gerard D. Wright suggested that it appears to be at least two distinctive classes of superbugs: the first enclosing well-known pathogens, many of which belong to the same genera, and even species, to that of the common human microbiota. These bacteria have undergone acquisition of AR genes, and frequently display increased virulence as well [77,82]. The foregoing class is thus characterized by bacterial pathogens, such as methicillin-resistant *S. aureus* (MRSA), and multidrug resistant *E. coli*. Minding that the former is usually sub-classified as either community associated MRSA (CA-MRSA), or health-care (hospital) associated MRSA (HA-MRSA), according to the epidemiological provenance of the infection, respectively [83]. The second class, however, encloses opportunistic pathogens, frequently of environmental origin, usually taking advantage of enfeebled or immunocompromised hosts. Said bacteria often include well-known opportunists like *Acinetobacter baumannii*, *Burkholderia cepacia*, *P. aeruginosa*, and *Stenotrophomonas maltophilia* [77]. In order to

better understand the very nature behind AR, one should be acquainted with its origins, and how this widespread phenomenon came into existence.

AR has been proclaimed to be an ancient characteristic amongst bacteria, predating the very first anthropogenic discovery, and therapeutic use, of antibacterial compounds by millennia [84,85]. This argument reinforces the notion that the wide environmental dissemination of AR elements is inconsistent with their relatively recent emergence, advocating instead the hypothesis of a rich natural history of AR [84-86]. Such a bold and startling statement can be succinctly explained through the lens of microbial ecology. As ulteriorly mentioned, spatially organized microbial communities have originated as soon as 3.25 billion years ago [2]. Yet, bacteria have been around much longer than that, taking into account that their origins date as far as 3.8 billion years [86]. Bearing in mind that a great deal of antimicrobials used as therapeutic agents are produced by environmental microorganisms (e.g.: Actinomycetes) [87], and that the genetic divergence of antibiotic-producing gene clusters places the origins of the former natural products at least hundreds of millions of years ago [88], we can conclude that bacteria have been under direct or indirect exposure to naturally-occurring antibiotics during an equal time period [86]. Furthermore, antibiotic-producing microorganisms must attain mechanisms as to deflect the very activity of the antibacterial compounds they produce; otherwise they would succumb to these toxic compounds, along with all other susceptible microorganisms [89]. These genetically encoded mechanisms of deflection are nothing more than resistance genes, which can be later transferred to human pathogenic bacteria, as well as other bacteria present in the same niche, in addition to the possibility of also being a product of independent evolution underwent by the former [79,86]. Corroborating on the first notion, we have that genome analysis of antibiotic-producers has shown the presence of genes pertaining to the same functional families as those that confer resistance to current populations of human bacterial pathogens [79,89]. Indeed, several functional metagenomic studies have proven that AR genes are disseminated throughout any studied microbial ecosystem, being able to confer resistance upon their exchange and further expression in a heterologous host [79,89,90]. Such is the case with the soil [89,90] and the human gut resistome [91] (the collective sum of all genes that confer direct or indirect resistance to antibiotics [77]). This fact pinpoints these environments as potential natural reservoirs for AR genes. These reports have been considered to be consistent with the emergence of AR in nosocomial environments, as well as the vastly described widespread dissemination of AR genes throughout characteristic microbial niches (e.g.: the soil) [84,89,90]. It has also been further predicted that novel antibiotics will only select for preexistent resistance genes, harbored for millennia within the resistome of these natural reservoirs [84].

As it can be ascertained from the previous paragraph, AR is far from a novel trait amongst bacteria. Still, the diversity of resistant bacteria, the global widespread of resistance, and the multidrug resistance present in single bacterial taxa has been certainly unprecedented and escalating in the past decades [92]. Regardless of the primordial origins of AR genes, the development and ensuing selection for generations of antibiotic-resistant bacteria, as well as their wide dispersion in microbial populations throughout the biosphere, are resultant from decades of ceaseless selection pressures superimposed by humans on the latter, through means of a brazen use, misuse, and abuse of antibiotic applications [93]. Strident examples of the formerly stated acts of carelessness are those of antibiotic prophylactic practices in non-clinical settings such as veterinary medicine, livestock production, animal husbandry, agriculture, and aquaculture, in addition to medical malpractice, unsupervised self-medication and their use in household cleaning products [92-95]. These actions don't come as a natural process, but instead as a man-made corruption of nature itself, which eventually led to the selection of antibiotic resistant clones as opposed to susceptible ones all across the biosphere, thus portraying the classical Darwinian

evolutionary processes of selection and survival [93].

Just as cleverly epitomized by Stuart B. Levy and Bonnie Marshall [80], selection for resistance can be construed as an equation with two main variables: the antibiotic, that inhibits susceptible organisms and selects those that are resistant; and the resistance-conferring gene present in microorganisms that were selected by the antimicrobial drug. In this fashion, drug resistance only comes forth when the two variables appear simultaneously in a given environment or host, which may in due course lead to a human health issue [80]. However, selection for resistance reaches way farther than that: if mediated by the expressed phenotype of a particular resistant variant in a bacterial clone, it not only selects the bacterial clone itself, but also all genes pertaining to the bacterial clone's genome and all the mobile genetic elements (e.g.: plasmids) and vectors contained in the former, as well as the genes contained within these elements and vectors themselves [79]. As a result, bacteria displaying an antibiotic resistant phenotype – along with the resistance genes under selection and the phenotype they confer – diffuse this genetic cohort as long as there are continuous antimicrobial selection pressures, to further exacerbate and extend the problem across other hosts and environments [80]. One must understand that the repercussions of the drug selection process can vary according to the geographical scale and quantitative density at which the antibiotic agent is being employed. If entire populations – whether they are humans, animals or plants – are imposing selective forces driven by the treatment with the same class of antibiotic, susceptible strains will have very little advantage within this niche, and as such, resistant strains will become the most apt to proliferate [80]. Thus resulting in a serious ecological imbalance that culminates in the emergence of an environmental pool of resistance genes within populations [80,96]. Furthermore, formulating on the density at which antibiotics are employed, it comes as extremely important to underline the effects of sub-lethal antibiotic concentrations on the resistance selective process, since these concentrations are the ones generally operating at natural environment scales, being either from direct anthropogenic pollution of the latter; those inherently generated by antibiotic-producing microorganisms; or the sub-lethal antibiotic concentrations present in the body compartments of humans or animals during the extent of therapeutic administration [97]. Since carrying an AR gene – either chromosomally or plasmid encoded – has implications in the fitness cost of the bacteria, its only logical to assume that if there is no selective pressure for AR in a given environment (i.e.: there are no antibiotic producers nor man-made antibiotic contamination), the toll of carrying such gene casts a competitive disadvantage to the bacteria in question, as opposed to susceptible bacteria which don't bear such gene, and consequently have no fitness cost associated. Nevertheless, this is not the case with our biosphere, since, like previously outlined, antibiotics are present at diverse concentrations throughout most microbial environments, hence providing a steadfast selection and maintenance pressure for resistant bacterial populations [93]. Several publications on the topic pertinent to the aftereffects of sub-lethal concentrations of antibiotics ultimately reckon that selection of antibiotic resistant bacteria occurs at exceedingly low antibiotic concentrations [97-100]. On a particularly prominent study [97], selective pressures imposed by three classes of antibiotics broadly used in clinical practice – aminoglycosides, fluoroquinololes and tetracyclines – with concentrations down to few hundred-fold below the minimal inhibitory concentration (MIC) of susceptible bacteria, could select and enrich a specific niche, for resistant ones. This could be partially explained by the fact that resistant bacteria have indeed a competitive advantage at all concentrations of a given antibiotic at which the susceptible clones' growth reduction is larger than the fitness cost of resistance [100], minding that even at sub-inhibitory concentrations, the antibiotic still exerts a burden on the susceptible clones.

Our discussion regarding the selection for antibiotic resistant bacteria wouldn't be thorough if we didn't approach how the multidrug resistance (MDR) phenomenon came into existence. Even though the

exact evolutionary mechanisms that led to the selection of MDR bacteria remain somewhat puzzling, a few explanations have shed some light on the subject. One intriguing aspect believed to be at the origins of MDR is that the continuous administration of a single antibiotic, selects for bacteria that are resistant to several other antibiotics, in addition to that particular one [92,101]. This occurrence suggests the presence of different resistance-conferring genes on the same mobile genetic elements, including, but not limited to, transposons and plasmids. Intriguingly, bacteria that already attained resistance to one antibiotic seem to earn competitive advantage by recruiting additional resistance genes from neighboring bacteria that share the same environmental niche [80]. Following this train of thought, it should be mentioned that a study which sampled the resistome of soil-dwelling bacteria, found on an average basis, that a sole bacterium displayed resistance to 7 to 8 different antibiotics [89]. Accordingly, combinatorial resistance has been proposed to be a standard phenotype amongst environmental bacteria [77]. This further ascertains the presence of MDR in opportunistic pathogens of environmental origin, such as the earlier cited *A. baumannii* and *P. aeruginosa*. It has also been further suggested that sub-inhibitory concentrations of antibiotics as well as multiple antibiotic combinations – frequently used in prophylactic and therapeutic clinical practice – might bring about the emergence and propagation of novel multidrug resistant bacterial pathogens via selection for resistance [102]. Notwithstanding all the abovementioned justifications, one of the simplest explanations for witnessing a never-ending increase in MDR, is that a single molecular mechanism, encoded by a single gene, may confer resistance to more than one antibiotic [103]. One such known resistance-conferring mechanism is that of multidrug efflux pumps.

Dwelling on the molecular mechanisms that confer resistance, and analogously to VFs, they present themselves as the product of resistance genes' translation into protein. These mechanisms have been thought to originate from naturally occurring antibiotic producing microbes [87,89], which as highlighted previously, were developed as a deflection tactic so that the former producers wouldn't suffer the same fate as the susceptible microorganisms neighboring them. Moreover, an antibiotic compound can only induce bacterial growth inhibition, or cell death, upon successful interaction with its target. In order for this to become possible the antibiotic must recognize the target, and the concentration of the latter compound must be enough as to achieve successful performance. Additionally, in furtherance of interacting with their targets, these compounds sometimes need to cross the bacterial cell envelope as well as be further activated by bacterial enzymes [79]. Therefore, the previous deflection mechanisms must fend off the modes of action by which antibiotic compounds operate, through elaborate strategies such as modification of the target (e.g.: through mutation or chemical modification); reducing the concentration antibiotic that can access the target (e.g.: through decreased permeability or active efflux); chemically modifying the antibiotic compound; or even protecting the target from the actions of the former compounds [79,104]. Ultimately leading to acquired resistance to the antibiotic targeting the bacterium in question. Said molecular mechanisms of deflection (i.e.: resistance determinants), congruently with the fact of having been extensively studied, can be classified in numerous ways. The most common being the classification according to which antibiotic they confer resistance to, and/or their functional role within the bacterium [93]. Focusing solely on the functional role of the resistant determinants, instead of the specific antibiotic they confer resistance to, and according to the Resfams database developed by Molly K. Gibson and associates [105], some of the most commonly depicted AR protein families can be succinctly classified as: (i) acetyltransferases; (ii) antibiotic inactivation enzymes; (iii) β-lactamases; (iv) gene modulating resistance proteins; (v) glycopeptide resistance proteins; (vi) multidrug efflux pumps; (vii) nucleotidyltransferases; (viii) phosphotransferases; (ix) quinolone resistance proteins; (x) rRNA methyltransferases; and (xi) target protection proteins [105]. Without entering into too much detail, the present dissertation shall disclose ever so briefly how these mechanisms operate, while focusing to a

greater extent on (two particular families of) multidrug efflux pumps.

We can see as an example that acetyltransferases, nucleotidyltransferases, and phosphotransferases confer high levels of resistance by chemically modifying the antibiotic compound – through means of transferring or switching an acetyl group, a nucleotide, or a phosphate group, respectively – ultimately rendering it unable to successfully interact with its target. On the other hand, β-lactamases and antibiotic inactivation proteins actively degrade this compound, making use of enzymatic machineries to achieve such goal [104]. From the former list one can argue that β-lactamases have been the most widely studied, as well as the ones with greater historical relevance since the very first use of antibiotics, with the disclosure of penicillinase (a β-lactamase that degrades penicillin) in 1940, just 12 years after the discovery of its target antibiotic [76,106]. As for function, β-lactamases operate through hydrolysis of the β-lactam ring (a ring comprised of four atoms, present in all β-lactam antibiotics), thus deactivating the drug's antibacterial properties [107]. Moving on to other mechanisms of resistance, ribosomal RNA (rRNA) methyltransferases for instance, methylate specific amino-acid residues in various bacterial rRNA subunits, conferring resistance to a wide range of drugs that are rendered unable of recognizing this site. Such is the case with the erythromycin ribosome methylase (*erm*) gene family, whose genes' products main function is to methylate 16S rRNA and alter the drug-binding site of macrolides, lincosamides and streptogramins [104,108]. The same can be analogously said for more loosely classified mechanisms like target protection proteins, glycopeptide resistance proteins, and quinolone resistance proteins, which also confer resistance by either chemically modifying the antibiotic's target – usually the DNA, DNA-associated proteins, or the ribosome – or by interfering with the antibiotic-target interaction. One such example of the latter mechanism is that of the *qnr* genes that encode for pentapeptide repeat proteins (PRPs) [104]. These PRPs act by binding to and protecting topoisomerase IV (a DNA-associated protein responsible for unlinking DNA following its replication) as well as DNA gyrase (a DNA-associated enzyme that relieves strain during the unwound of double-stranded DNA) from the bactericidal action of quinolones [104], being that a relatively recent study further conveys that PRPs interact with the topoisomerase-quinolone complex after the binding of the antibiotic, resulting in the release of the quinolone [109]. Gene modulating resistance comes as a broad classification of various AR determinants, which may play a more indirect part in resistance. The term "gene modulating" refers to their role as modulators of AR gene expression, bringing into mind that the vast majority of these so-called modulators can act at the gene transcription, or translation level. Moreover, their actions can be occasionally induced by the presence of antibiotic compounds, sometimes at sub-lethal concentrations [110]. A common example of these determinants is the two-component system VanR/VanS of enterococci, which directly controls the expression of genes that mediate resistance to vancomycin [110]. This two-component signal transduction system is composed of a membrane-dwelling histidine kinase (VanS), and a response regulator present in the cytoplasm (VanR) that acts as a transcriptional activator of many vancomycin resistance genes, including *vanA*, *vanH*, and *vanX*, amongst others [110,111]. The expression of other AR mechanisms, for instance multidrug efflux pumps, is not usually regulated by two-component systems, although some exceptions have been reported, namely that of the AdeABC pump in *A. baumannii*, which is in due term regulated by the AdeR/AdeS two-component system [110,112].

Even though AR can be attained through a plentiful array of molecular mechanisms, particularly those depicted so far, resistance due to active efflux of drugs poses as a mechanism of paramount importance when it comes to the AR thematic, since a single class of multidrug efflux pumps can confer resistance to various antibiotics, bestowing a MDR phenotype to the bacterium expressing these efflux mechanisms [113]. Illustrating on the common mode of action that ultimately confers AR to bacteria that

comprise these efflux pumps, one can perceive that the latter mechanisms operate by extruding antibiotic compounds from within the cytoplasmatic milieu onto the extracellular space, after these compounds had been previously internalized, in an active or passive way, by the bacterium. In this fashion, the efflux pump-coding bacterium never reaches inhibitory intracellular concentrations of antibiotics, and as such, becomes resistant to the latter. Albeit these resistance determinants were firstly characterized in *E. coli* as being plasmid-borne and assumed to be acquired through means of HGT [114], just like the majority of resistance determinant-encoding genes reported until then, it soon became clear that these mechanisms weren't an exclusive trait of bacteria, nor were they restricted to being plasmid-encoded, minding that genes encoding for these determinants were later found encoded in the chromosome of other bacteria and also in the chromosomes of archaea, and even eukaryotic organisms, as duly reviewed by José L. Martínez and colleagues [115]. It also comes as important to outline that the presence of multidrug efflux pumps is not circumscribed to antibiotic producers, just as the fact that, in bacteria, most multidrug efflux pumps encoding genes described insofar have been found to be enclosed within the chromosome, exhibiting a well-conserved structure, along with a firmly regulated expression [115,116]. In addition to these characteristics, the expression of multidrug efflux pumps is not confined to bacteria that dwell amidst environments with high antibiotic selective pressures, since most of these pumps not only extrude antibiotics, but also toxic compounds resultant from human industrial activity (e.g.: organic solvents derived from petroleum); other antimicrobials, like those produced by plants; bacterial signaling molecules (e.g.: quorum sensing molecules); and even heavy metals, originated from human environmental pollution or naturally-occurring in our planet's crust [115]. Thus reinforcing either the non-selective nature portrayed by these efflux pumps, as well as their wide dissemination in bacteria pertaining to non-clinical environments, and partially explaining why the presence of a single class of these mechanisms can confer resistance to a myriad of different antibiotics. It has been further suggested that the aforementioned characteristics advocate that the core function of these systems mightn't have risen with the purpose of extruding antibiotics used in clinical-practice, considering all the other roles portrayed by said mechanisms that pose as relevant to bacterial behavior and survival within their natural ecosystems, including, but not limited to, detoxification from intracellular metabolites, maintenance of cell homeostasis and intercellular signal trafficking [115]. Notwithstanding the broader functions exerted by multidrug efflux pumps, and rather directing ourselves towards their categorization, one can be acquainted with the five families of multidrug efflux pumps that have been described thus far, according to their structural configuration, number of regions that span the membrane(s), sources of energy, and substrates: (i) the ATP-binding cassette (ABC) super-family; (ii) the multidrug and toxic compound extrusion (MATE) family; (iii) the major facilitator super-family (MFS); (iv) the resistance-nodulation-division (RND) super-family; and (v) the small multidrug resistance (SMR) family (which is a subgroup of the drug/metabolite transporter super-family – DMT)[113,117]. Only the ABC super-family and the MFS super-family of multidrug efflux pumps shall be disclosed as mechanistic examples of AR determinants. Multidrug efflux pumps pertaining to the ABC super-family export (or import) a wide variety of substances, driven by the energy discharged from ATP hydrolysis, being the latter process the core feature of this super-family, which stands as the biggest of all paralogous families of proteins [113,118]. The minimal, yet sufficient, structural organization required of a functional ABC efflux pump comprises four domains: two ATP-binding domains, and two membrane-spanning permease domains, taking into account that extrusion systems are either characterized by homodimeric, or heterodimeric organization, in which one nucleotide binding domain is fused with one transmembrane permease domain, being usually called half-transporters [118]. As an example of an ABC multidrug efflux pump, we have the MacB macrolide exporter of *E. coli*, that exists as a dimer and contains the

four characteristic domains from all ABC proteins [119]. In the former bacterium, the MacB efflux pump, which spans the inner membrane, works together with MacA, a membrane fusion protein from the periplasmic space, and TolC, an outer membrane channel protein. MacA couples ATP hydrolysis from MacB ATPasic activity by promoting a closed ATP-bound state on the latter, with the transport of substrates across the outer membrane through TolC, thus establishing a physical link between the ABC efflux pump (MacB), and the outer membrane channel (TolC) [119]. This complex is usually referred to as the MacAB-TolC tripartite efflux system that in such a way manages to span both inner and outer membranes of *E. coli* [119]. Furthermore, as a second relevant example of an ABC multidrug efflux pump, MsbA in *E. coli*, a Lipid-A flippase, plays a role in the biogenesis of the outer membrane by being responsible for the transport of lipid A across the inner membrane [118]. Resistance conferred by MsbA in *E. coli* has yet to be documented. However, MsbA present in *Lactococcus lactis* has been reported to confer resistance to erythromycin, an antibiotic from the macrolide class, and showed high levels of a DNA dye and ethidium export [120]. Conversely to ABC efflux pumps' mechanistic mode of action, MFS efflux pumps are secondary active, or passive transporters, only capable of extruding low molecular weight solutes, via ion gradient or solute osmosis, respectively [121,122]. Even so, both MFS and ABC super-families of multidrug efflux pumps are recognized as the most prevalent drug efflux systems in bacteria [118]. *E. coli*, for instance, has 10% of its genome coding for efflux pumps of the MFS-type [123]. Members of the MFS were believed to uptake sugars as their primary function [121], since these proteins' super-family comprises either uniporters (transport of a single substrate in favor of the gradient), symporters (transport of a solute and an ion in favor of the gradient), or antiporters (transport of a single substrate against the gradient) [122], bearing in mind that MFS efflux pumps fall into the antiporter category [123]. Yet, nowadays it's well established that the substrates mobilized by MFS transporters are discrete small molecules such as amino acids, antibiotics, nucleic acids, sugars, and various metabolites [122]. These efflux pumps are of extreme interest to researchers since they're easily able to confer multidrug resistance in critical bacterial pathogens, being well-documented candidates for therapeutic modulation [122,123]. Concerning their structure, and as a rule of thumb, MFS efflux pumps have either 12 or 14 transmembrane spanning segments (α-helices) [122]. According to Indrika Ranaweera et. al. [122], and taking a 12 segment efflux pump into consideration, we can further see that these segments organize themselves in two bundles, symmetric in structure but asymmetric in function, named N-terminal (transmembrane spanning segments 1-6) and C-terminal (transmembrane spanning segments 7-12) domains, respectively. Creating a large central aqueous cavity formed by the elements of these two bundles as an end result [122]. Likewise to ABC-type pumps, in Gram-negative bacteria, MFS efflux pumps can also be part of multicomponent systems, like the well-studied EmrAB-TolC complex of *E. coli* [123]. Similarly to the ABC-type tripartite system mentioned above, the EmrAB-TolC is comprised by EmrB (the MFS efflux pump), EmrA (the periplasmic adaptor protein) and TolC (outer membrane channel), forming a contiguous efflux system that allows direct export of substrates from the cytoplasm, straight to the extracellular space, thus achieving the same function as the abovementioned one [124].

So far this present subsection has been discussing the emergence of AR determinants; their wide dissemination throughout the biosphere; the evolutionary processes by which they're selected; and the most notorious molecular mechanisms of AR. But perhaps the most important topic regarding AR was purposefully left for last. Such topic is that of Horizontal gene transfer (HGT). Outlining a review penned by Cheryl P. Andam and colleagues [125], HGT can be briefly defined as the exchange of genetic material between two cells that do not partake in an ancestor-descendant relationship, and contrarily to parent-to-offspring inheritance, acknowledges almost no taxonomic boundaries amongst bacteria. The

genetic material exchanged might be comprised of gene fragments; whole genes; operons (cluster of genes controlled by a sole promoter – DNA region that initiates transcription); superoperons (operons that encode complex metabolic pathways); plasmids, and even entire chromosomes [125]. Conjugation, transformation and transduction are consistently referred to as the three most well studied methods of HGT in bacteria [126]. Conjugation is the exchange of genetic matter between a donor and a recipient cell, making use of a conjugation pilus, requiring physical contact between the former cells. Conjugative plasmids are commonly the mediators of such process, although conjugative transposons are also known to induce the conjugation process [39]. Transformation is construed as the uptake of "free" exogenous DNA from the surrounding environment, an illustrious feature of competent bacteria – competence being defined as the innate ability of a given cell to uptake "free" DNA. Finally, transduction can be seen as the consignment of genetic material as a consequence of bacteriophage predation, assuming that genetic material from another host (i.e.: bacterium) has been previously integrated into the bacteriophage's genome [126]. After the exchange of genetic material, the recipient cell now possesses additional molecular information, which can confer a new phenotypic trait, like resistance to a given antibiotic. Thus, this inherent ability displayed by microbes (e.g.: bacterial pathogens), might pose – in some particular instances – as a great threat to human health [125]. Taking plasmids as a relevant example, one can reckon the easiness through which they are spread across heterogeneous bacterial populations [39], and the fact that such plasmids often code for AR genes [127], and even virulence-conferring ones, the latter being often associated with cooperative traits amongst the preceding microbes [128]. Another menace originates from the rapid emergence, mutation, and positive epistasis (interaction between genes) of AR genes present on disease-causing organisms [129] that might display a MDR phenotype. The genetic information enclosed therein can be later transmitted to other bacteria, such as those that dwell within the human body (e.g.: the human gut microbiota [130]), generating possible shifts between mutualism and parasitism [4,128], and exasperating the AR phenomenon even further. Following the preceding example, one already knows that the gastrointestinal tract has been revealed to be a reservoir of AR genes for some time now [91], owing to an optimal blend of conditions that promote the emergence and spreading of AR genes amongst bacterial populations, namely high cell density [130]. Another factor is that of HGT, by which resistant mutualistic bacteria may bestow resistance genes upon transient bacteria that don't reside within the gut for extensive periods of time (e.g.: pathogens) [130]. For instance, a study conducted on human volunteers showed that resistance to ampicillin and sulfonamide was transferred by means of a conjugative plasmid, from one resident *E. coli* strain, to another *E. coli* strain that had been administered [130,131], advocating such hypothesis. This phenomenon has also been proclaimed to purportedly occur the other way around – from transient pathogens to our microbiota. One such example is that of a study conducted during an outbreak of *Enterobacter cloacae*, which showed that conjugation of a carbapenem resistance-conferring plasmid probably happened between the offending pathogen and other Enterobacteriaceae present in a patient's gut microbiota, further spreading to other subjects [130,132]. Another example is that of a reported *E. coli* strain that donated a plasmid, encoding for ampicillin resistance, to another *E. coli* strain in an infant's gut [130,133]. The authors of such study concluded that the selective pressure required as to increase the donor cell density – thus favoring the transference of the resistance-encoding plasmid to the receiving strain – was provided following the administration of ampicillin, in order to treat the patient for a urinary tract infection [130,133]. With the given examples, one can undoubtedly ascertain the preponderance that HGT bears when it comes to the known dissemination of AR genes, that often seem to find a way to walk hand-in-hand with virulent bacterial lifestyles.

1.4 Co-selection Hypothesis

It cannot be overemphasized that one of the major concerns regarding pathogenic bacteria is their increasing ability to resist antibiotic treatment. Referring back to the previous section, this overbearing phenomenon has severely hindered our ability to clinically approach infectious bacterial diseases worldwide [77-79]. Moreover, these ailments are often induced by MDR pathogens, like MRSA [83], or multidrug resistant Tuberculosis (caused by the *Mycobacterium tuberculosis* bacterium) [134], known to reemerge from time to time in a multitude of successive outbreaks throughout the world. Further reiterating on virulence and AR as deterring factors of human survival, and in spite of the fact that both are extensively scrutinized themes of foremost importance to microbiology, and medical practice in general, there have been very few instances where they are conjointly addressed, since they have, for a long time, been outwardly seen as unlinked phenomena [135].

In 2002, José L. Martínez and Fernando Baquero outlined this fact in a comprehensive review [135], where they asked if an evolutionary relationship between AR and bacterial virulence is in fact taking place, focusing on the selective pressures imposed by modern medical practices, by means of an ever-increasing antibiotic usage, as the main driving force operating on the co-selection of AR and bacterial virulence. As one already knows, the dawn of therapeutic antibiotics initially precluded the spread, and quite possibly the evolution of pathogenic bacteria, with the aftermost costly price of AR emergence [135]. Indeed, from an ecological standpoint, one should also recollect that both infection-states and antibiotic therapies generate extremely effective evolutionary bottlenecks [30,93], that in due course prune bacterial diversity, only enabling host colonization to a very selective subset of highly specialized bacteria able to withstand these stringent conditions [135], bearing in mind the selective pressures exerted by either the host immune system, antibiotic therapy, or even the two combined. Martínez and Baquero further asserted that virulence and AR can be construed as similar biological mechanisms of adaptation, which have been selected through a lengthy evolutionary course, encompassing countless generations, as to grant bacterial survival under the outlined ecologically stressful conditions [135]. Additionally, even though AR and virulence determinants were developed on disparate timescales [136], from a biological standpoint they still share common characteristics nonetheless, such as (i) both being required by bacteria in order to survive under inauspicious environments [30,80,135,136]; (ii) both being mainly acquired through means of HGT, from other bacteria [31,127,128]; (iii) direct or indirect interplay between both resistance and virulence-conferring determinants, with special reference to the part played by adhesion mechanisms and multidrug efflux pumps in biofilm-producing pathogens [53,136,137]; and (iv) being sometimes controlled by regulatory systems, which activate or repress the expression of genetic complexes that encompass both types of determinants [136,138].

Heeding the role HGT plays in the co-selection phenomenon, one can also realize that, in point of fact, the transference of mobile genetic material has been proposed to be the quintessential genetic process by which the dissemination, and subsequent co-selection of both virulence and resistance genes occurs [136]. If we acknowledge a conjugative plasmid as a germane example, coding both for VFs and AR determinants, one knows that it could swiftly spread across heterogeneous bacterial communities [39], such as those that inhabit the human gastrointestinal tract. Hence, further reminiscing on the fact that in bacteria, HGT mechanisms seem to be confined by few, or none, taxonomic barriers [125], one can picture that this conjugative plasmid becomes widespread within the aforementioned communities, bestowing the encircled bacteria with a genotype that codes both for resistance and virulence. Elaborating on this scenario, and ascertaining the preponderance of conjugation on the

co-selection of these determinants, one should also be acquainted with the fact that this process has been reported to trigger the bacterial SOS response – bacterial stress response resultant from an abnormal increase of single-strand DNA (ssDNA), which is often an indicator of DNA damage – known to control both DNA repair and gene recombination [139]. On a side note, it ought to be mentioned that bacterial SOS response has been established to be commonly induced by the presence of antibiotics, such as quinolones which directly inflict damage upon the DNA [136], as well as the fact that sub-inhibitory concentrations of the former compounds have also been reported to heighten bacterial transcription, and quite possibly enhance, or favor, the acquisition of favorable traits, through the very process of gene transfer [99,140]. These low concentrations can be easily achieved due to the misuse of antibiotic drugs, by means of environmental pollution or prophylactic administration, as accordingly stated in the previous subsection. In such a way, HGT is hereby prompted through diverse processes like conjugation, integration and transposition. And even if antibiotics are correctly administered in a clinical setting, the development of antibiotic gradients is renown to exist, either within a human body, or in an environmental scope [99].

As to better contextualize the part played by bacterial SOS response in the co-selection process, one can fathom that ssDNA is produced by several HGT mechanisms, for instance, conjugation [139]. As such, when conjugation takes place (or when an antibiotic damages bacterial DNA), a sudden escalation of ssDNA levels follows, being readily detected by the recipient bacterium, which ensues a SOS response, ultimately leading to gene cassette (mobile genetic element that contains a gene – or genes – and a recombination site) rearrangements [139]. Gene cassettes are known to often include AR genes, as well as being usually incorporated within integrative genetic elements called integrons [141]. These elements can be defined as genetic units that contain determinants coding for site-specific recombination systems, capable of gene capture and mobilization, such as those comprised by the foregoing gene cassettes [141]. In this fashion, once a bacterium initiates a SOS response, it up-regulates integron integrases (the enzymes responsible for the integration of genetic material), that consecutively capture and incorporate gene cassettes, leading to further rearrangements of the former genetic complexes [139]. The preceding mechanism can, quite plausibly, drive the acquisition of a MDR phenotype, at least through the agency of two different events: if one postulates multiple gene cassettes coding for resistance to different antibiotics being successively incorporated every-time this mechanism takes place; or a sole gene cassette that encodes, by itself, for a multidrug resistance determinant (e.g.: efflux pumps) [103,142].

Although integrons are not mobile elements by their own nature, they are known to exist in plasmids and transposons [142], elements that bestow them with mobility. Accordingly, one can reckon that any given virulence-coding plasmid, containing integrons in addition to the latter determinants, can easily incorporate multiple AR genes cassettes, via integrase recombination. Along these lines, this hypothetical plasmid, becomes a vector for the dissemination of both AR and VFs – occurrence which has already been depicted in the literature [143]. Nevertheless this event can happen the other way around: a resistance-conferring plasmid can incorporate genetic islands coding for VFs, such as the VCR (*Vibrio cholerae* repeated sequences) cassettes, which are prototypical examples of composite integrons encasing virulence determinants [135,144]. Furthermore, research on VCR clusters has established that the genetic organization of these elements is akin to that of gene cassettes coding for AR known to be present in integrons [135,144]. Also, the provenance of VCR islands has been presumed to have risen from integrase mediated processes, which further advocates on the similar role VCR islands play in gene capture, when comparing to that of integrons in the procurement of AR genes [135]. Notwithstanding all these instances, an integron can also be transposon-borne, like briefly mentioned above [142]. And if one recalls that transposons are mobile elements that can readily change their position within a given

genome (e.g.: between other mobile genetic elements, like plasmids, and the bacterial chromosome), one can also acknowledge the suppositional aggravation these genetic elements might exert when taking the conjoined capture of virulence and resistance-conferring genes as well as their later dispersal to other bacteria into account – although knowledge concerning the relationship between virulence and resistance determinants in the same transposon is still lacking, albeit for a few peculiar examples [136]. By the very nature of the processes described thus far, it becomes quite clear that if the same genetic element codes for both AR genes and VFs, then, co-selection of both types of determinants will surely ensue, i.e, if a given bacterial population carries these coupled determinants, the selection for virulent traits (e.g.: through means of selective pressure imposed by the host immune system) will consequently select for AR determinants; and analogously, antibiotic selective pressure (e.g.: enforced by antibiotic therapy) will consequently select for VFs [135]. Now, imagining an alternative theoretical scenario, for instance, concerning the human gut microbiota, in which its residing bacterial population detains a well-conserved chromosomal gene coding for resistance to a certain antibiotic, resultant from long-term antibiotic consumption by the host. Amidst this milieu, there might be a transient alien pathogen as well, containing a plasmid coding for virulence. If this pathogen, through HGT, shares his plasmid with one bacterium within the gut microbiota, and this bacterium further shares said plasmid with the rest of the community, this hypothetical community now codes for both virulence and AR, rendering the bacteria resident therein as potential pathogens. This situation may also happen the other way around, i.e., the gut microbiota encompassing a resistance-conferring plasmid and then sharing it with the transient pathogen that chromosomally encodes for virulence, which in due course becomes resistant to a given antibiotic, being now able to infect other hosts, while portraying both virulent and resistant traits. One must understand however, that this imaginary situation does not correspond to co-selection per se, but rather to co-representation of both types of determinants. Yet, there is another setting on which co-representation of both determinants might arise. Since microbial communities, together with the environment that surrounds them, can be seen as purported functional evolutionary units [1], if in the same microbiome there happens to be bacteria whose genome encodes for VFs, whilst other bacteria encode for AR determinants, when analyzing said microbiome as a whole (e.g.: metagenomic analysis), a co-representation of both types of determinants might present itself, even though they are not being encoded in the same bacterium. Nonetheless, acknowledging that these bacterial communities can be regarded to function as a whole, one can define this type of associations as being indeed of co-representative nature, being able to later prove themselves as co-selective if the aforesaid mechanisms of HGT come into play amidst the addressed communities.

In addition, conjugative plasmids have also been reported to induce biofilm production [145], which itself plays a crucial role in HGT. The very matrix-like structure of biofilms promotes HGT, especially by means of conjugation. Indeed, the very process of conjugation also ends up stimulating the production of biofilm structures, given the high cell density and closeness of the foregoing bacterial cells [136], thus resulting in a feedback loop of sorts. Other HGT processes like transformation also seem to be required for biofilm formation and further stabilization [136,146]. Biofilms should not be seen as virulence-conferring complexes per se, minding that a great deal of bacterial populations have been widely depicted as being found in association with environmental surfaces, throughout numerous ecosystems, in this type of complex that usually encloses several species [145]. However, some determinants that lead to biofilm production are known to be VFs, like type IV pili, fibronectin-binding proteins, and a type IV secretion system which has been established to be a contributor in the formation of the former complexes, mediating cell-to-cell contact, and by this way, intervening in DNA transfer, respectively [46,53,136,147]. Still, when in an infectious setting, biofilms are matrix-complexes

of undeniable importance, conferring outstanding advantages to the pathogens that produce them, outlining their direct implication in several austere diseases, such as chronic bronchitis, cystic fibrosis, endocarditis, kidney stones, and osteomyelitis [135,148]. When in this type of setting, biofilms allow bacterial pathogens to subvert host immunological responses [148], and even confer an indirect resistance to the actions of several antibiotics, bearing in mind that bacteria growing in biofilms are more resilient than those that lead a planktonic lifestyle [135], since the actions of the former antibiotics do not affect bacteria inhabiting the inner layers of biofilms. Moreover, multidrug efflux pumps have also been acknowledged as one of the main mechanisms that grant AR to biofilm-producing bacteria [137]. Efflux systems have been thoroughly involved in bacterial signaling – i.e.: quorum sensing – regulation, and it is well established that quorum sensing controls the expression of numerous VFs, along with biofilm differentiation [137,149]. Evidence concerning the role of multidrug efflux pumps in biofilm resistance has been found in several bacteria like *P. aeruginosa* and *E. coli* [137]. Several studies have reported that, in *E.coli* for instance, biofilms detain higher AR in comparison to planktonic cells, and that expression of several multidrug efflux pumps was found to increase in biofilms [137,150], just like the genes coding for the AcrAB-TolC multidrug efflux pump system, which have been proclaimed to be up-regulated under the formerly mentioned conditions, in addition to exposure to several antibiotics [137]. Without intending to deviate from the subject at hand, one should also be mindful of the existence of a *P. aeruginosa* two-component regulatory system (CbrAB), implicated in biofilm formation, virulence, and AR [138]. In this regulatory system, CbrA is a sensor kinase, propounded to modulate biofilm formation, cytotoxicity, and even swarming motility through CbrB regulatory responses, and may likely modulate AR independently [138]. Now, if one acknowledges all the abovementioned characteristics that are inherent to biofilms and the processes involved therein, one can therefore ascertain that there is in fact not only a positive feedback loop between HGT processes – asserting on conjugation as the most relevant process – and biofilm formation [136], as there also appears to be an interplay between AR determinants and VFs, working together as to reach the same goal. These instances might very likely favor the transmission of AR and VFs genes collectively, especially in the presence of antibiotic selective pressures [136]. Minding that antibiotics might act as to select for bacteria capable of producing biofilms, thusly broadening the prevalence of reputed infective lifestyles [135], as well as the interrelationships, common gene regulators, and co-representation of both types of determinants.

As a concluding remark, one can realize that in a world with great availability of antibiotics, and in some cases, unregulated administration, bacteria and microbiomes are subject to different levels of antibiotic selective pressure. In this context, we can envisage that for some pathogens, in order to survive and colonize the host, it is not enough to code for VFs, if antibiotics are present. That is to say that AR can be viewed as being part of the pathogen's strategy. Or, from another point of view, under antibiotic selective pressure, selection of mobile genetic elements coding both for resistance and virulent traits might occur, resulting in their dissemination within human bacterial communities, such as the human gut microbiota. Even if these determinants do not find themselves within the same mobile genetic element, or in the same bacterium for that matter, the co-representation of the former determinants might also reinforce on the notion of co-selection, given the presence of antibiotic selection pressures.

1.5 Metagenomics and Bioinformatics

Metagenomics came into existence as an astonishing brave new field of study, following the post-genomic era. It can be construed as contrastive with genomics, in the way that the latter rather concerns itself with the study of individual genomes from single organisms, whereas metagenomics addresses all the genomes, and further genes, from all the microbial representatives enclosed within the given sequenced sample collectively – recollecting on the notion that designates the aforementioned collection of genes as a metagenome. This rather recent field, together with the ever-increasing availability of new high-throughput novelties regarding sequencing technologies, has been fomenting an explosion of data, that once scrutinized, aids like never before, in the expansion of our knowledge pertaining to ecological, metabolic, and physiological processes taking place amidst the "hidden world" of environmental microbial communities. Indeed, like Delmont et. al. gracefully emphasized [151], this fairly new field of study has helped us unravel what the "black box" of environmental microbial communities contains. For example, an incommensurable abundance of novel genes, which can in due course aid us in our endeavors concerning matters of indisputable importance, like the discovery of new compounds bearing pharmacological interest, and even towards a better understanding of environmental processes, such as those relating to agronomical settings, climate change and degradation of pollutants [151]. Yet, one of the most important contributions that rose with early advents of environmental gene sequencing is undeniably the ability to better determine the phylogenetic diversity encompassed within a given environmental sample [152], since it has been long established that more than 99% of all microbial life found in nature cannot be consistently cultivated through standard procedures [153]. Thus, culture-independent methodologies, including the comparative analysis of small ribosomal RNA subunits (mainly 16S rRNA), previously established by Olsen, Pace and colleagues in 1986 [154], laid the foundations for further innovations in light of metagenomic analysis, carving the path for more recent approaches, that are currently commonplace. Particularly PCR (Polymerase Chain Reaction) and whole-genome shotgun oriented sequencing techniques, that ultimately bestow researchers with a collection of predominantly unbiased data, representing the vast majority of genes pertaining to all microbial representatives of the given sampled communities [155].

Although contrastive with a sole genome, one might, quite presumably, acknowledge a metagenome as a discrete unit of genetic information, for several reasons. Firstly, one should keep in mind that bacteria are social organisms, mostly living in well established communities, that together, might function as a purported biological system, relying on the full set of genes that the microbiome comprises. Within their naturally occurring communities, bacteria typically form close cooperative loops resulting in indirect benefit to all species involved [156], being able of complementing each other, in order to reconstitute complete biochemical pathways and metabolic functions. Thusly, it comes as no surprise that those metabolic exchanges and symbiotic biochemical interactions are found to be ubiquitous amidst microbial communities [157,158]. Secondly, HGT promotes interrelationships between bacterial species, compelling them to cooperate [128], and consequently avoiding the emergence of cheaters (bacteria that benefit from the cooperation of other bacteria, without contributing themselves) within the microbiota. Thirdly, by addressing a metagenome, one is granted access to the repertoire of genes involved in adaptation to the environment, as well as cooperation [159], of the bacterial communities represented within the sequenced sample, mostly due to the fact that most of these traits are often encoded in mobile genetic elements [159], and thus can be shared by different, eventually unrelated bacteria. For these reasons, one can conceive a metagenome as a representative unit of all bacterial communities' genomes encompassed within the environmental sample that has been sequenced.

Still appertaining to the subject at hand, one can also accede to the fact that mining for genes in metagenomes comes as a trustworthy way to access the selective pressures a given bacterial population is being subject to, as well as the co-selection, or co-representation, of genetic traits comprised by the microbiome as a whole. The accomplishment of mining for genes in metagenomes, together with the unceasing submission of new data, and the consistent increase of public access to the copiousness of biological information depicted in latter, throughout several databases, annotation and analysis platforms [160-162], might aid microbial ecologists in their research endeavors, by providing answers to biological questions, as well as supporting the evaluation and development of new hypotheses [151]. Notwithstanding the cited databases, the present dissertation, for all its purposes, would like to especially distinguish the MetaGenomics RAST (Rapid Annotation using Subsystems Technology) server (MG-RAST) [160], which has underwent an amazing expansion ever since its debut in 2008, either in the quantity and diversity of metagenomic data and even whole projects upholded within said database [163], with a great deal of them being publicly available; as in the broad variety of automatically pre-processed files that undergo a file formating pipeline as to provide the user with sequence quality assessment and annotation with reference to numerous renown databases, along with a user-friendly post-annotation platform for further analysis and visualization of metagenomic data [164]. In addition to all these features, one should also mention that all metagenomic data gathered within this database is conveniently reachable through its application program interface (API) [165].

Shifting one's attention to the sum of microorganisms that share our body space, also known as the human microbiota [7], one might as well be acquainted with the fact that it has become a highly focused theme of current research concerning microbial ecology [8,166]. Some examples of groundbreaking research conducted over the past few years on the human microbiota have assuredly been directed towards the niche that comprehends the highest distribution of bacteria within our body – the human gut. With special emphasis on a few prominent topics, including but not limited to, host-microbiota relationships [5]; the ecological and evolutionary features regarding the former milieu [4]; and with the bacterial diversity subsumed therein, along with its subsequent genetic and metabolic diversity [8,12,13,167]. As to underline a pertinent example, one can be apprised with a rather illustrious study, authored by Tanya Yatsunenko and colleagues [167], that delved on the genetic and metabolic diversity depicted by gut microbiomes amongst human populations having contrastive cultural, sanitary, dietary, and socio-economic lifestyles [167]. This study probed bacterial diversity through means of 16S rRNA gene analysis, in fecal samples donated by 531 individuals (aged 0 to 70 years), and also analyzed the complete set of gut metagenomes from a subset of 110 subjects (aged 0 to 53 years), belonging to a cohort that enclosed healthy children and adults spanning from two regions of the Venezuelan Amazon, three provinces of Malawi, and four cosmopolitan areas of the USA [167]. One of several findings provided by the appliance of the foregoing methodology, showed that the first three years of an individual's life presented core features of the gut microbiota's functional maturation, characteristics that were identified as being shared by all three populations [167]. And that during this period the bacterial phylogenetic diversity present in the human gut increases gradually, almost linearly, confirming once again that the human gut microbiota begins establishing itself right upon birth, undergoing maturation during the first years of life. Like it had already been reported in, at least, one previous study that also made use of 16S rRNA gene analysis [168]. Furthermore, after this 3-year time span, the diversity of bacterial taxa seems to reach a succeeding plateau, later stabilizing throughout adulthood [167]. Another interesting result included unmistakable discrepancies in bacterial communities and functional gene repertoires between individuals from the USA and those native to the other two countries, being these distinctive traits evident from early infancy, throughout to adulthood [167]. One might also be further cognizant

of the fact that the conclusions drawn from this particular study, raise a whole new assortment of very interesting questions and prepositions. Whereas the research conducted by these authors vehemently suggests that the human gut microbiota must be reconsidered when evaluating various human features, such as development, nutritional status, and physiological disparities; additionally concerning itself with the impact of westernization on the human gut microbiome, and how these changes might potentially mediate the suite of pathological states [167]; this dissertation independently oughts to ask whether different levels of access to putative medical care, and therefore pharmaceutical drugs, intrinsic to different and contrastive human populations, also render different grades of AR determinants and VFs acquisition, by the gut microbiome of said populations.

The microbial diversity and community dynamics rendered by the human gut microbiota is constitutional to a vast myriad of physiological and metabolic processes that ultimately contribute to the endowment of a healthy-state to its host [166]. Researchers have been aware of the connection between the mammalian immune system and gut microbial communities for many years [166], but even so, contemporary research continues to unravel the intricacies of said relationship [169]. However, our general health is prone to a whole spectrum of intertwined and concomitant competencies exerted by the gut microbiota, that reach far beyond such affair [166,170]. Many diseases, regularly complex and dysbiotic in nature [21,22], have been continuously correlated with changes to the microbiota and its microbiome [166], unfolding, along these lines, how intrinsically connected one is with its gut microbiota, together with the role it plays in health and disease [15-19,166,170]. In addition to the previous statement, if one further reiterates on the third subsection of this dissertation's introduction, one can recollect that opportunistic bacterial pathogens often arise from environmental settings [77], and that environmental microbiomes pose as abundant reservoirs of AR genes [84,89,105], along with human gut microbiomes [91], sometimes even overlapping and sharing the same resistance determinants [90]. Indeed, the presence of AR genes amidst the human gut microbiome, and all the implications this phenomenon bears, has been extensively studied by several research groups worldwide, with special reference to the one led by Gautam Dantas [91,105,171-174]. Some of the most recent findings, collectively gathered by his team, have concluded that environmental factors shared within families shape resistome development in healthy infants, as early as a few weeks after birth, even without exposure to antibiotics [171,173]; that the gut microbiome of Amerindian communities, with no known previous contact with westernization, nor pharmaceutical-grade antibiotics, were shown to carry AR genes, syntenic with mobile genetic elements [172]; and the presence of core AR genes in low-income human settings, that cross environmental boundaries, bearing possible associations with HGT phenomena [174], just to name a few. The latter settings can very well add up to a co-representation of both resistance and virulence determinants encoded by potentially opportunistic pathogens dwelling amidst our gastrointestinal tract – for all the aforementioned reasons stated in the previous subsection –, boldly asserting on the ever-growing evidence that seems to show that AR determinants are in fact widespread throughout human gut microbiomes, even in healthy newborn individuals [171,173]. Indeed, metagenomic studies have already shown us that there is, as a matter of fact, shared presence of pathogenic species of bacteria, such as those currently defying medical practice, and AR genes, amongst microbial communities dwelling in the environment, especially those under direct anthropogenic influence, for instance, manured soils [175], and wastewater treatment plants [176,177]. Nevertheless, to the knowledge of all group members involved in the very project which has led to the writing of this dissertation, there are no scientific records on the evolutionary dynamics ruling the epidemiology of resistance and virulent bacterial determinants collectively, in any reported biome, despite the preexistent awareness the scientific community possesses regarding the success and speed of

bacterial adaptation, concerning AR together with the emergence of multi-resistant pathogenic bacteria.

Mining for genes, or protein sequences, amongst metagenomic data commonly relates to a central paradigm of most bioinformatics and computational biology studies – the inference of sequence similarity through the virtue of computational search algorithms. As William R. Pearson expounded in an original and enlightening review [178], the step that appertains to sequence similarity search of homologous sequences is usually one of the first, and most informative milestones in any analysis referring to the characterization of newly described sequences [178]. Briefly recounting the annals of how sequence similarity searching algorithms came to be, one should know that it all began with the heuristic approach devised by Needleman and Wunsch in 1970 [179], which first introduced a calculation method reliant on a substitution matrix (a matrix that describes the rate at which a character – nucleotide or amino acid – present in a given sequence, changes to another character over time), allowing the conferral of a score (i.e.: similarity score) as the end result of the alignment procedure applied to the totality of the two sequences being compared. This alignment methodology is commonly referred to as the global alignment algorithm, and it's still presently used for optimal global sequence alignment, especially when the end quality of the global alignment precedes all other requirements. Meanwhile, numerous other heuristic algorithms were suggested, but unfortunately these were either devoid of biological significance, or uninterpretable [180]. Even so, in 1974 Sellers developed of a true metric measure, allowing the calculation of the disparity between two given sequences, representing in this fashion the minimum number of mutations – insertions, deletions or substitutions – required in order to convert one sequence into another [181]. Yet, a long-lasting success decisively came, with the contrivance of the local alignment algorithm, by Smith and Waterman in 1981 [180]. The Smith-Waterman algorithm differs from the one designed by Needleman and Wunsch, almost a decade earlier, in the instance that instead of comparing the totality of the whole two sequences being aligned, it compares several segments of all possible lengths from these sequences, thus optimizing their similarity score as an end result. It should be mentioned that for all current bioinformatics purposes, one does not usually implement the Smith-Waterman algorithm per se, being acquainted with the fact that nowadays, there are better alternatives bearing improved scalability [182], and accuracy [183].

There are, at least, two strong catalytic factors that often lead researchers to choose sequence homology search programs based on local alignments, in deterrence of global alignment ones. Firstly, one should know that obtaining correct alignments in regions that reveal low similarity between distantly related biological sequences is a computationally challenging task. Mainly because mutational events add too much undercurrent information, over the evolutionary timescale that separates both sequences being compared, to concede a substantially accurate comparison of the depicted regions [184]. As such, local alignment algorithms avert these regions entirely, rather focusing on those that share evolutionarily conserved similarity cues [184]. Secondly, there is a statistically sound model for local alignments [185,186], enabling in such a way, a certain degree of reliability for optimal local alignments, further allowing the calculation of better scores for the latter. These scores follow an extreme value distribution, when taking the alignment of unrelated sequences into account. The preceding feature concedes the creation of an expectation value (E-value) – based on the P-value correction applied to multiple testing – for the optimal local alignment between the two sequences being compared [185], which comes as a common statistical estimate of many homology search programs. Simply put, this E-value is an estimation of how often two unrelated sequences would generate an optimal local alignment, having by chance, a greater or equal score to the one that is being observed [185], given the total number of sequences being considered in a multiple testing scenario (e.g.: database size) [187]. Very low E-values indicate that the two sequences in question might be homologous, therefore they might share a common

ancestor, and might even, possibly, have similar structures and functions. However, the quality of a match also depends on other criteria, such as the total length of the alignment produced, and similarity percentage. For this reason, such statistical measures should also be taken into account when considering the significance of a match. Even so, and despite the fact that a somewhat ordinary unspoken rule seems to be that two protein sequences can be reported as homologous if they share more than 30% identity over their entire length, this criterion might miss many homologies that would be detectable if one regarded statistical estimates, like E-values, instead [178]. Albeit the fact that alignments sharing 30% identity, and bearing at least 100 or more residues in length, are practically always statistically significant, many homologies are readily detected with E-values $<$ 1e-10, that are nowhere near a 30% identity threshold. Hence, E-values should be regarded as much more useful for inferring homology than identity percentages [178]. Moreover, the implementation of the commonly used 30% identity criteria as to pinpoint homology between two sequences, harshly underestimates the total number of detectable homologs by sequence similarity algorithms between very distantly related species (e.g.: yeast and human), seeing that protein homologs might actually share less than 20% identity [178]. Furthermore, since E-values are reliant on the total number of multiple comparisons being made – that is to say, the size of the database –, alignment scores inferred upon searching a bigger database will yield less significance than the exact same scores inferred upon searching a smaller database instead. Nonetheless, the previous statement does not imply that the resulting alignments prove homology in one context whereas in the other they do not. If an alignment is significant – thus providing evidence for homology – in a search with a database that has a smaller number of entries, the same alignment can also be indicative for homology in a bigger database, however it might not bear significance, since there are more sequences being taken into account upon the E-value's calculation, increasing the number of alignments that could yield a significant score by chance [178].

As of today, the sheer number of bioinformatics tools and heuristic homology search programs that, through modifications of the original Smith-Waterman algorithm, provide local sequence similarity search is tremendous. A few examples might encompass older software like FASTA [188], SSEARCH [189], and the renown BLAST (Basic Local Alignment Search Tool) [184], along with an expanded version, that provides DNA-to-protein (BLASTX), protein-to-protein (BLASTP), and even multiple sequence alignments (PSI-BLAST) [187]. Newer algorithms, and subsequent programs, include BLAT [190], USEARCH [191], HMMER3 [192], and the latterly DIAMOND [193]. These latest programs, although depicting a faster performance than BLAST [190-193], or even having a different statistical reasoning behind its algorithms [192], still lack the statistical reliability and high sensitivity intrinsic to BLAST, which remains after all these years, as the "golden" standard regarding local alignment tools [193]. Despite this fact, out of all the illustrated programs above, BLAST still falls short on accuracy when taking even older, but more precise, programs into account. For instance, SSEARCH implements the rigorously accurate Smith-Waterman algorithm as modified by Gotoh [182], however its performance speed is orders of magnitude slower than other programs [189], if not the slowest. Therefore, BLAST can be relatively seen as the most reliable choice regarding the compromise between speed and sensitivity, as well as its extensive validation throughout the concerning literature, still being recognized to this date, as the one possessing the best ratio with respect to the prior variables.

Attending to this dissertation's objectives, only homology inferred from protein-to-protein alignments shall be addressed, for the following reasons correspondingly: (i) Since there are only 20 amino acid residues that partake in the synthesis of proteins, but 64 codons in the genetic code, a certain degree of degeneracy and redundancy arises [194]. Even though some amino acids are known to be encoded by unique codons (e.g.: methionine and tryptophan), others have as much as

six codons encoding them (e.g.: arginine, leucine and serine) [194]. This degeneracy portrayed by the genetic code is the prime factor responsible for the existence of synonymous mutations, also known as silent mutations (substitution of one nucleotide for another in a protein-coding gene, in such a manner that the complementary codon encodes for the same amino acid, rendering the resulting amino acid sequence unmodified) [194]. For the preceding reason, and also bearing in mind (ii) different codon preferences, depicted by different organisms (i.e.: codon usage bias [195]), when aligning sequences one might find a much lower level of identity considering nucleotide-to-nucleotide comparisons, than when taking protein-to-protein sequence comparisons into consideration. (iii) Following William R. Pearson's reasoning [178], one can fathom that similarity searches encompassing protein sequences are much more sensitive than those enclosing DNA, given that nucleotide-to-nucleotide alignments barely discern homology if the concerning sequences have last shared common ancestry more than 200 ~ 400 million years ago, whilst protein-to-protein alignments promptly detect homologous sequences that underwent divergence more than 2.5 billion years ago [178]. (iv) Likewise, alignment statistics pertaining to DNA comparisons are less accurate than protein-to-protein ones [178]. While protein-to-protein alignments bearing E-values $<$ 1e-03 can assuredly be used as to ascertain homology, E-values $<$ 1e-06 pertaining to DNA alignments are recurrently reported to happen by chance [178]. As such, similarity searches enclosing protein sequences (e.g.: BLASTP) should be favored, for the simple reason of being five to tenfold more sensitive, and statistically accurate, than similarity searches comprising DNA sequences [178].

1.6 Objectives

As to conclude, one can acknowledge that the rationale for this dissertation concerns itself with (i) bioinformatical mining of publicly available metagenomic data stored in a well-established database [160], as to ascertain the diversity of protein homologues coding for AR determinants and VFs, enclosed within environmental and human gut metagenomes sampled worldwide [151,167], by making use of a renown protein sequence homology search algorithm [187], as well as original subsequent filtering algorithms; (ii) with understanding if these determinants are in fact co-represented in metagenomes (with the aid of simple linear regressions), and if the co-representation pattern remains the same throughout three different human populations depicted in the human gut dataset; (iii) with understanding the single and co-associated pattern portrayed by these determinants throughout the ages of the individuals from whom the human gut metagenomes were retrieved; and finally, with (iv) statistical analysis of the magnitude of associations portrayed by both types determinants encompassed therein. Both linear regressions and statistical approaches shall be accordingly discussed in the methods section.

2. Methods

2.1 Metagenomic datasets

This project's human gut query cohort enclosed 110 publicly available metagenomes pertaining to individuals issuing from different regions of Venezuela (21 metagenomes), Malawi (23 metagenomes), and the USA (66 metagenomes), as well as comprising a broad age span (0.05 to 53 years), as previously depicted in the article authored by Yatsunenko et. al. [167]. The environmental query cohort comprised 64 previously selected, and publicly available metagenomes, belonging to 12 different biomes, including acid mines drainage biofilms; Antarctic aquatic environments; chicken cecum; coral atolls; cow rumen; deep oceans; human faeces; mouse cecum; oceans; Phosphorous removing sludge; sediments; and soils. All of which had already been antecedently mentioned in Delmont et. al.'s article [151].

The MG-RAST accession numbers belonging to the 110 human gut metagenomes, were extracted from the metadata file ("jobs table") that had been formerly downloaded from the respective project's MG-RAST dedicated webpage (`http://metagenomics.anl.gov/metagenomics.cgi?page=MetagenomeProject&project=98`). However, the MG-RAST accession numbers referring to the 64 environmental metagenomes had to be extracted from the article's appendix, since they are coming from a collection of selected metagenomes, spanning from different and independent projects. It should also be mentioned that although Delmont's team [151] report using a dataset comprised of 77 metagenomes, there are only 70 MG-RAST accession numbers present in the article's appendix, of which only 64 are publicly available.

The metagenomes pertaining to the human gut dataset were downloaded on the 3rd of April 2015, and the metagenomes belonging to the environmental dataset were downloaded on the 17th of November 2015, from the MG-RAST database (`http://metagenomics.anl.gov/`) [160], under FASTA format, respectively. The download process made use of successive calls to the MG-RAST API [165], automated by means of a Z-Shell script that used the respective MG-RAST accession numbers available from the aforesaid sources as arguments. Each FASTA file comprised protein-coding sequences, retrieved from the MG-RAST file-formatting pipeline (550.cluster.aa90.faa files), clustered at 90% homology, containing only non-redundant translated sequences. According to the MG-RAST team [164], these protein-coding sequences were identified with FragGeneScan [196], and further clustered at 90% identity with Cd-hit [197]. As such, each FASTA file contains the protein sequences of one representative from each generated cluster, along with all the singleton sequences that were left unclustered. Thus they portray the protein diversity enclosed within each metagenome.

2.2 BLASTP, VFDB, Resfams and file processing

For every metagenome present in our query cohorts, a BLASTP [187] search was achieved against the 2012 version of the Virulence Factor database (VFDB) of bacterial virulence factors protein families [198], and the Resfams AR Proteins database of bacterial antibiotic resistance protein families [105]. The BLAST+ executables package (ncbi-blast-2.2.31+ version) was downloaded on the 17th of November 2015 from the NCBI website (`ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+`). The VFDB was downloaded on the 11th of November 2013 from its respective website (`http://www.mgc.ac.cn/Vfs/v3index.htm`), enclosing a total of 31 functionally-classified FASTA files of bacterial virulence factor protein sub-families (see Table 2.1), and the Resfams AR Proteins database was downloaded on the 29th of January 2016, from Dantas Lab's website (`http://www.`

dantaslab.org/resfams), encompassing a total of 123 functionally-classified FASTA files of bacterial antibiotic resistance protein sub-families (see Table 2.2). A previous approach made use of the Antibiotic Resistance Genes Database (ARDB) [199], but several hindrances concerning it's sub-classification by functional antibiotic resistance protein families made us discard the possibility of using such database.

All FASTA files pertaining to these databases were successively formatted by a Z-Shell script that made use of the *makeblastdb* application as to produce BLAST databases from FASTA files, with the non-default parameter for the creation of BLAST protein databases (*-dbtype prot*). Every BLASTP search was automated by way of a Z-Shell script that used both the protein-coding clustered FASTA files from our query cohorts' chosen metagenomes, along with the VFDB and Resfams databases' FASTA files as arguments. Each BLASTP search was performed with non-default parameters for an E-value cut-off of 1e-15 (*-evalue .000000000000001*), and the number 6 tabular output file format (*-outfmt "6"*). The total number of BLASTP searches, and subsequent output files, for the chosen query cohort enclosing human gut metagenomes against both databases was of: 110 (human gut metagenomes) * 31 (VFDB files) + 110 (human gut metagenomes) * 123 (Resfams files) = 16940 outfiles; and for the chosen query cohort enclosing environmental metagenomes against both databases was of: 64 (environmental metagenomes) * 31 (VFDB files) + 64 (environmental metagenomes) * 123 (Resfams files) = 9856 outfiles.

Table 2.1: VFDB FASTA files classified by their mechanism and protein family function.

VF mechanism (FASTA file)	Protein Family Function
Chaperone/Usher pathway	Adhesion & Invasion
Extracellular-nucleation-precipitation pathway	Adhesion & Invasion
Type IV pili	Adhesion & Invasion
Sortase-assembled pili	Adhesion & Invasion
Flagella	Adhesion & Invasion
Autotransporters	Adhesion & Invasion
Fibronectin-binding proteins	Adhesion & Invasion
Fibronogen-binding proteins	Adhesion & Invasion
Collagen-binding proteins	Adhesion & Invasion
Others	Adhesion & Invasion
Type II secretion systems	Secretion Systems & effectors
Type III secretion systems & effectors	Secretion Systems & effectors
Type IV secretion systems & effectors	Secretion Systems & effectors
Type V secretion systems	Secretion Systems & effectors
Type VI secretion systems & effectors	Secretion Systems & effectors
Type VII secretion systems & effectors	Secretion Systems & effectors
alpha-PFT	Toxin
beta-PFT	Toxin
Superantigens/superantigen-like proteins	Toxin
Surface-acting enzymes	Toxin
ADP-ribosyltransferase	Toxin
Glucosyltransferase	Toxin
Guanylate/Adenylate cyclase	Toxin

Continued on next page

Table 2.1 – continued from previous page

VF mechanism (FASTA file)	Protein Family Function
Deaminase	Toxin
RNA N-glycosidase	Toxin
Metalloprotease	Toxin
DNase I / genotoxin	Toxin
Intracellular PFT	Toxin
Siderophore-mediated Iron Uptake	Iron Acquisition
Heme-mediated Iron Uptake	Iron Acquisition
Transferrin and Lactoferrin-mediated Iron Uptake	Iron Acquisition

Table 2.2: Resfams AR Proteins FASTA files classified by their mechanism and protein family function.

AR mechanism (FASTA file)	Protein Family Function
ABC Antibiotic Efflux Pump	ABC Transporter
macA	ABC Transporter
macB	ABC Transporter
msbA	ABC Transporter
tolC	ABC Transporter
AAC3	Acetyltransferase
AAC3-Ia	Acetyltransferase
AAC6-Ia	Acetyltransferase
AAC6-Ib	Acetyltransferase
AAC6-II	Acetyltransferase
Chloramphenicol Acetyltransferase CAT	Acetyltransferase
TE inactivation	Antibiotic Inactivation
TetX	Antibiotic Inactivation
BCII	Beta-Lactamase
BJP	Beta-Lactamase
BlaB	Beta-Lactamase
CARB-PSE	Beta-Lactamase
CblA	Beta-Lactamase
CepA	Beta-Lactamase
CfxA	Beta-Lactamase
ClassA	Beta-Lactamase
ClassB	Beta-Lactamase
ClassC-AmpC	Beta-Lactamase
ClassD	Beta-Lactamase
CMY-LAT-MOX-ACT-MIR-FOX	Beta-Lactamase
CTXM	Beta-Lactamase
DHA	Beta-Lactamase
DIM-GIM-SIM	Beta-Lactamase
Exo	Beta-Lactamase
GES	Beta-Lactamase

Continued on next page

Table 2.2 – continued from previous page

AR mechanism (FASTA file)	Protein Family Function
GOB	Beta-Lactamase
IMP	Beta-Lactamase
IND	Beta-Lactamase
KHM	Beta-Lactamase
KPC	Beta-Lactamase
L1	Beta-Lactamase
LRA	Beta-Lactamase
MoxA	Beta-Lactamase
NDM-CcrA	Beta-Lactamase
PC1	Beta-Lactamase
Sfh	Beta-Lactamase
SHV-LEN	Beta-Lactamase
SME	Beta-Lactamase
SPM	Beta-Lactamase
Subclass B1	Beta-Lactamase
Subclass B2	Beta-Lactamase
Subclass B3	Beta-Lactamase
TEM	Beta-Lactamase
VEB-PER	Beta-Lactamase
VIM	Beta-Lactamase
baeR	Gene Modulating Resistance
baeS	Gene Modulating Resistance
blaI	Gene Modulating Resistance
blaR1	Gene Modulating Resistance
marA	Gene Modulating Resistance
mecR1	Gene Modulating Resistance
mprF	Gene Modulating Resistance
phoQ	Gene Modulating Resistance
ramA	Gene Modulating Resistance
robA	Gene Modulating Resistance
romA	Gene Modulating Resistance
soxR	Gene Modulating Resistance
vanR	Gene Modulating Resistance
vanS	Gene Modulating Resistance
vanA	Glycopeptide Resistance
vanB	Glycopeptide Resistance
vanC	Glycopeptide Resistance
vanD	Glycopeptide Resistance
vanH	Glycopeptide Resistance
vanT	Glycopeptide Resistance
vanW	Glycopeptide Resistance
vanX	Glycopeptide Resistance

Continued on next page

Table 2.2 – continued from previous page

AR mechanism (FASTA file)	Protein Family Function
vanY	Glycopeptide Resistance
vanZ	Glycopeptide Resistance
emrB	MFS Transporter
MFS Antibiotic Efflux Pump	MFS Transporter
norA	MFS Transporter
TetA-B	MFS Transporter
TetA-G	MFS Transporter
TetA	MFS Transporter
TetD	MFS Transporter
TetE	MFS Transporter
TetH-TetJ	MFS Transporter
Tetracycline Resistance MFS Efflux Pump	MFS Transporter
TetY	MFS Transporter
ANT2	Nucleotidyltransferase
ANT3	Nucleotidyltransferase
ANT4	Nucleotidyltransferase
ANT6	Nucleotidyltransferase
ANT9	Nucleotidyltransferase
ANT	Nucleotidyltransferase
Macrolide Glycosyltransferase	Other
adeR	Other Efflux
adeS	Other Efflux
Chloramphenicol Efflux Pump	Other Efflux
emrE	Other Efflux
APH3 double prime	Phosphotransferase
APH3	Phosphotransferase
APH3 prime	Phosphotransferase
APH6	Phosphotransferase
Chloramphenicol Phosphotransferase CPT	Phosphotransferase
Fluoroquinolone Resistant DNA Topoisomerase	Quinolone Resistance
Quinolone Resistance Protein *Qnr*	Quinolone Resistance
adeA-adeI	RND Antibiotic Efflux
adeB	RND Antibiotic Efflux
adeC-adeK-oprM	RND Antibiotic Efflux
MexA	RND Antibiotic Efflux
MexC	RND Antibiotic Efflux
MexE	RND Antibiotic Efflux
MexH	RND Antibiotic Efflux
MexW-MexI	RND Antibiotic Efflux
MexX	RND Antibiotic Efflux
RND Antibiotic Efflux Pump	RND Antibiotic Efflux
16S Ribosomal RNA Methyltransferase	rRNA Methyltransferase

Continued on next page

Table 2.2 – continued from previous page

AR mechanism (FASTA file)	Protein Family Function
ArmA	rRNA Methyltransferase
Cfr23 Ribosomal RNA Methyltransferase	rRNA Methyltransferase
Erm23S Ribosomal RNA Methyltransferase	rRNA Methyltransferase
Erm38	rRNA Methyltransferase
ErmA	rRNA Methyltransferase
ErmB	rRNA Methyltransferase
ErmC	rRNA Methyltransferase
TetM-TetW-TetO-TetS	Target Protection
Tetracycline Resistance Ribosomal Protection Protein	Target Protection

Next, a filtering algorithm was applied, under the form of an AWK script, as to screen the output files for alignments that fulfilled the requirements for a minimum of 60% sequence "homology" (percentage of the query that aligned with the subject more than 60%) over a 75% alignment overlap. All protein sequences that were unable to fulfil these criteria were accordingly discarded. Furthermore, for a given protein present in a metagenome used as a query, only the best hit (i.e.: alignment score) for that specific protein was retrieved from its respective output file, after being validated by the preceding filtering algorithm. Afterwards, duplicate hits for proteins that seemed to characterize both antibiotic resistance determinants as well as virulence factors – i.e., proteins which had, at least, aligned with both a protein sequence present in a Resfams database file, and a protein sequence present in a VFDB file – were removed, by virtue of an algorithm also implemented through a Z-Shell script. Hits (i.e.: alignments) that passed all filtering procedures described insofar were considered valid. This way we have taken into account the proteins of each metagenome that are homologues to either an AR or a VF determinant.

Two tables (data warehouses) enclosing all valid hit counts from the filtered output files, were consecutively created, for each metagenomic dataset, also through the usage of yet another Z-Shell script: one table for the hit counts resultant from the BLASTP search of the foregoing metagenomes against the Resfams database; and another one for the hit counts resultant from the BLASTP search of the foregoing metagenomes against the VFDB, respectively. The sums of all VFDB, and Resfams hits, were further filtered for uniqueness per metagenome, upon the tables' creation – a given protein sequence identifier, from its respective metagenome, was not represented more than once, amidst the total number of representatives. These two tables were then merged into one final table via a Python script, for both datasets, proceeding differently, when attending to the disparate natures of their respective metagenomes' metadata: for the human gut metagenomic cohort, both tables were merged along with the metagenomes' MG-RAST accession numbers, total count of protein sequences present in each metagenome, name-code of the metagenomes, together with the age, location and country of the individuals from which the metagenomes were gathered; whereas for the environmental metagenomes, both tables were merged along with the metagenomes' MG-RAST accession numbers, total count of protein sequences present in each metagenome, and the respective biome from which the metagenomes were gathered. The abovementioned filters and algorithms were implemented making use of UNIX scripting languages (GNU AWK version 4.0.1, and Z-Shell version 5.0.2), and the Python programming language (version 2.7.6), under a Linux environment. For a concise diagram depicting all the workflow described thus far please refer to Figure 2.1.

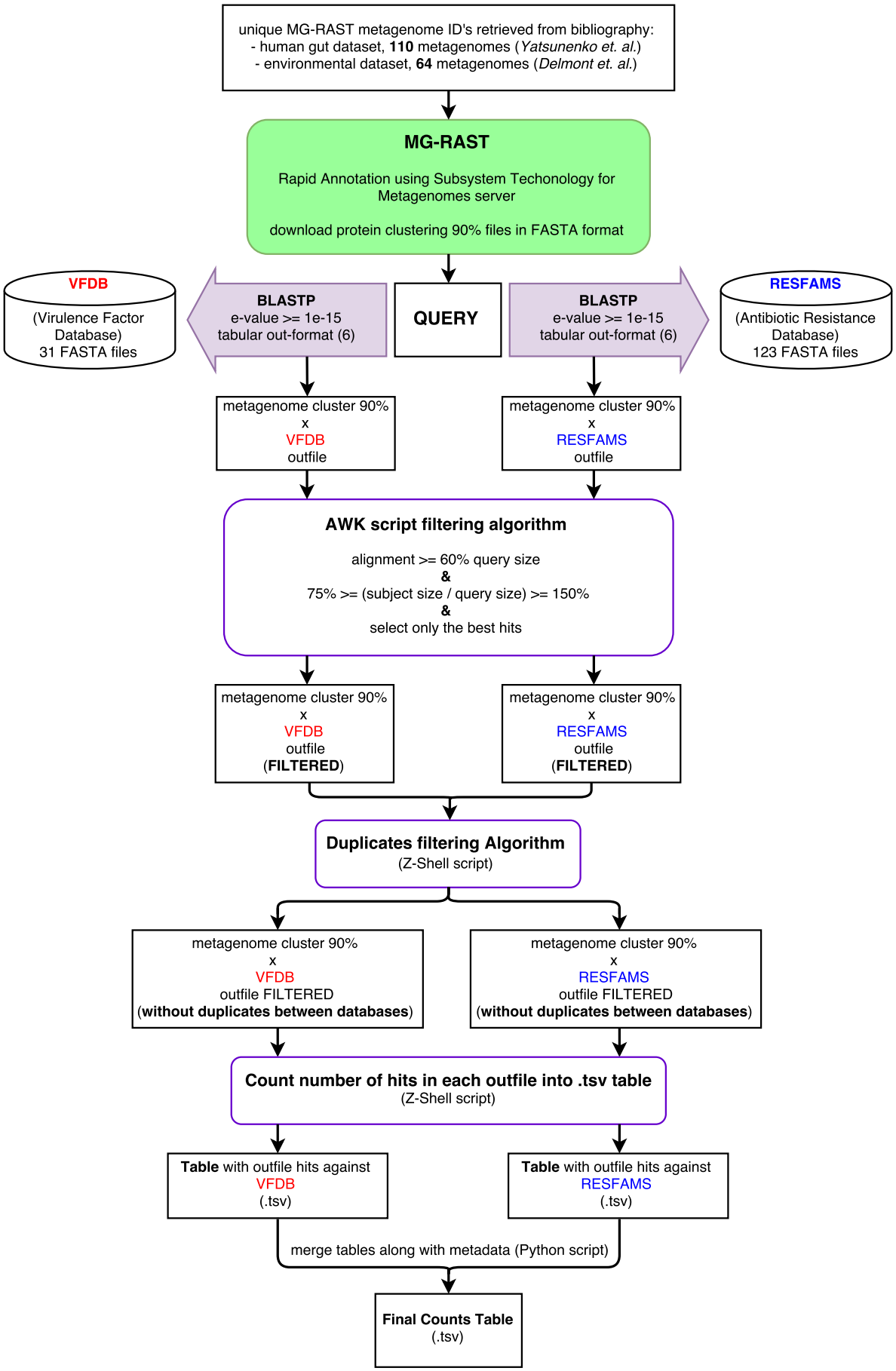


Figure 2.1: Flowchart of the implemented file-formatting workflow.

2.3 Linear Regressions and Statistical Analysis

The major aim of our work was to check whether a metagenome with more ARd than expected, had more or less VFd than expected – where ARd is the diversity number (hit counts) for homologues of antibiotic resistance protein families, and VFd is the diversity number for homologues of virulence factors protein families, respectively. One should expect that the number of ARd (or of VFd) increases with the metagenome's total number of protein sequences (its "size") according to the "law of diminishing returns". However, to simplify our analysis, and because there are many proteins in each functionally-characterized family of AR and VF determinants, we shall assume a linear relationship between ARd and metagenome size, whilst doing the same for VFd. Our results show that the assumption of linearity is reasonable. In order to study such relationship, we proceeded as follows. Given the assumption of linearity, we use the simple linear regression formula:

$$y_i = ax_i + b \qquad \text{(Equation 2.1)}$$

Where, following the least-squares approach, x is a variable believed to hold a linear relationship with the other variable y; a is the slope of the linear fit; and b is the y-intercept. Further along, and as to better interpret the linear relationship between ARd and VFd, these hit counts were divided by the slopes extracted from their linear fit with the total protein sequence count of their respective metagenomes, and further divided by the total protein sequence counts as well, respectively (see Equation 2.2 and 2.3. for the corresponding simplified formulae). This way ARd and VFd were plotted on the same scale, and were also accordingly standardized by the total number of protein sequences contained in their respective metagenomes. Moreover, we define α as the slope of the regression line of ARd on the size of the metagenomes, and β as the slope of the regression line of VFd on the size of the metagenomes:

$$ARd = \alpha.Size \qquad \text{(Equation 2.2)}$$

$$VFd = \beta.Size \qquad \text{(Equation 2.3)}$$

Thus, a metagenome i of a given size (*Size(i)*) is expected to have:

$$ARd(i) = \alpha.Size(i) \qquad \text{(Equation 2.4)}$$

antibiotic resistance protein families' homologues, and:

$$VFd(i) = \beta.Size(i) \qquad \text{(Equation 2.5)}$$

virulence factors protein families' homologues (to clarify the meaning of these equations, we draw Figure 2.2 with hypothetical examples). This is the case of metagenomes "1" and "3" in Figure 2.2. Naturally, some of the metagenomes do not match these predictions (the case of all metagenomes of Figure 2.2, with the exception of metagenomes "1" and "3"). As such, we ask what happens when these predictions are not met (Figure 2.3), being that different contexts can be conceived.

Suppose that a given virulence plasmid is very epidemic and spreads among several species of a bacterial community. After that, an incoming similar plasmid coding, for example, for antibiotic resistance, would have a certain degree of difficulty in order to spread and stabilize amidst the foregoing

community. The reason underlying this occurrence is that, being similar to each other, the two plasmids are incompatible, meaning that they are not stable in cells – one of them is lost during bacterial division. For this hypothetical and specific situation, the presence of certain virulence genes implied an absence of certain antibiotic resistance genes, and consequent gene products, respectively. In other words, this virulence plasmid would increase VFd but decrease the expected value of ARd for this metagenome, a trade-off rule: more genes of a certain kind implying less of another, albeit it might be the other way around.

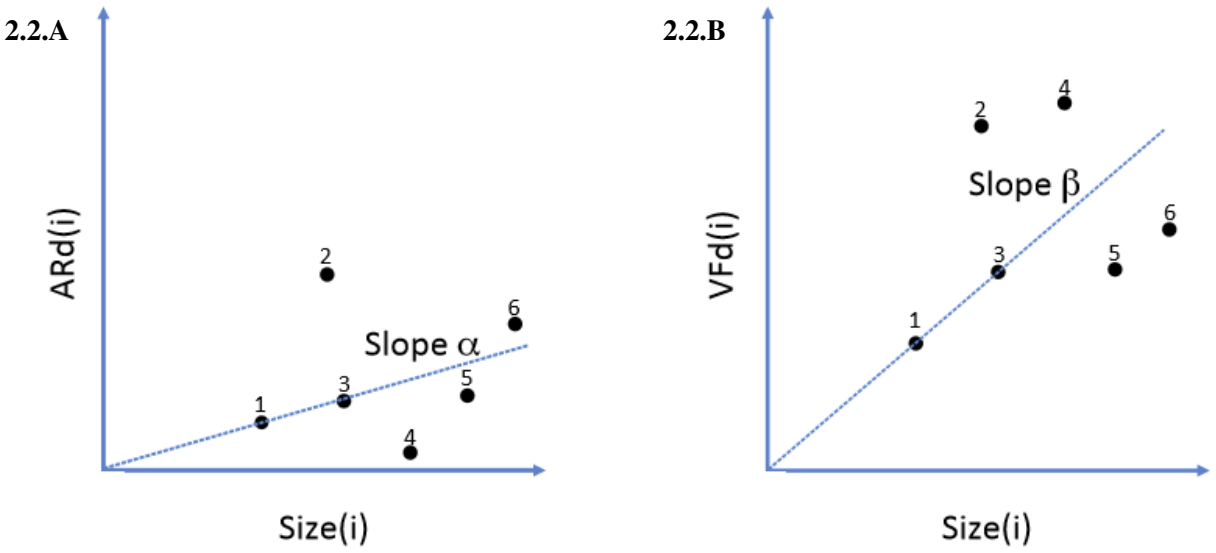


Figure 2.2: Schematic representation of ARd and VFd counts of six hypothetical metagenomes and their relationship with the size of each metagenome.
Slopes α and β are those of Equations 2.4 (A) and 2.5 (B), representing the slope of the regression line of ARd on sizes and of VFd on sizes, respectively. Metagenomes are numbered between "1" and "6". Metagenomes "1" and "3" fall on the line. Metagenome "2" has more ARd (diversity number for homologues of antibiotic resistance protein families) and more VFd (diversity number for homologues of virulence factors protein families) than expected, metagenome "5" has less ARd and less VFd than expected. Metagenome "4" has less ARd than expected, but more VFd than expected, and metagenome "6" an excess of ARd but a deficit of VFd (the opposite of metagenome "4"). Should we expect more instances like those of metagenomes "2" and "5" or like those of metagenomes "4" and "6"? Or all of them equally?

As a second hypothetical situation, one can suppose that a given bacterial genome, encompassed within a given sequenced human gut metagenome, confers virulence. As to fight this pathogen the host eventually had to undergo antibiotic treatment, thus selecting for antibiotic resistance genes. These antibiotic resistance genes may even be encoded by another bacterial genome (belonging to the same microbiome, and further metagenome). According to this scenario, the presence of a given VF in a metagenome implies selection of a certain AR gene, hence ultimately portraying co-representation of VF and AR determinants (and not a trade-off situation akin to the one before). This circumstance leads to an increase of both ARd and VFd, eventually driving to an excess of both.

Finally, we may also expect a third scenario: that there is no general rule for outsiders. Strictly speaking, some metagenomes with an excess of ARd may also have an excess of VFd (or deficit of both), but other metagenomes with an excess (or deficit) of ARd have a deficit (or excess) of VFd.

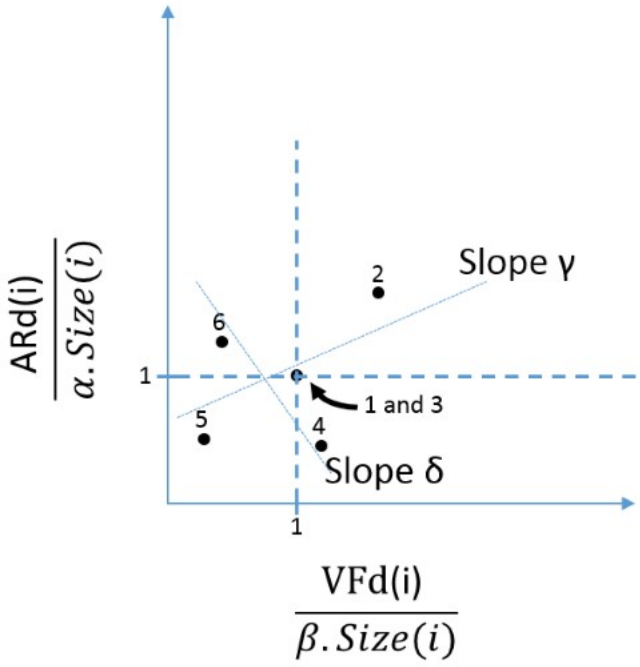


Figure 2.3: Relative expected ARd and VFd of outsider metagenomes.
Metagenomes "1" and "3" of Figure 2.2 would appear on coordinates (1,1). According to Figure 2.2, metagenome "2" has more ARd (diversity number for homologues of antibiotic resistance protein families) and more VFd (diversity number for homologues of virulence factors protein families) than expected, and metagenome "5" less ARd and VFd than expected. Metagenome "4" has less ARd than expected but more VFd than expected, whereas metagenome "6" has more ARd and less VFd than expected. If most metagenomes are like metagenomes "2" or "5", the regression line amongst metagenomes has a positive slope (represented here as slope γ). If most metagenomes are like "4" or "6", we should expect a negative slope (represented as slope δ). The possibility of no relationship between the two variables may also occur.

One can now realize that for a given metagenome *i*, it is expected that:

$$\frac{ARd(i)}{\alpha.Size(i)} = 1 \qquad \text{(Equation 2.6)}$$

and that:

$$\frac{VFd(i)}{\beta.Size(i)} = 1 \qquad \text{(Equation 2.7)}$$

Being that the expectation under the null hypothesis is that the values of the two sets fall under the (1,1) coordinates (see Figure 2.3). Should one expect these ratios to be simultaneously lower or higher than one, or, on the contrary, if one of them is greater than one, the other one should be lower than one? Perhaps, there might even be different forces ruling this relationship (see Figure 2.3).

As such, the Spearman rank correlation coefficient (Spearman's ρ, or r_s) was used to test the association between the standardized ARd and VFd. All possible associations between AR and VF determinants' hits for protein sub-families were also generated (n = 123 Resfams files * 31 VFDB files = 3813). As to control the expected proportion of rejected null hypotheses, and bearing in mind that this is a multiple comparison scenario, the Benjamini-Hochberg (BH) post-hoc procedure [200] was applied to the r_s P-values generated this way, ensuring a certain degree of correction over the false-discovery rate, without giving in to the stringency of a Bonferroni correction (see Figure 2.4).

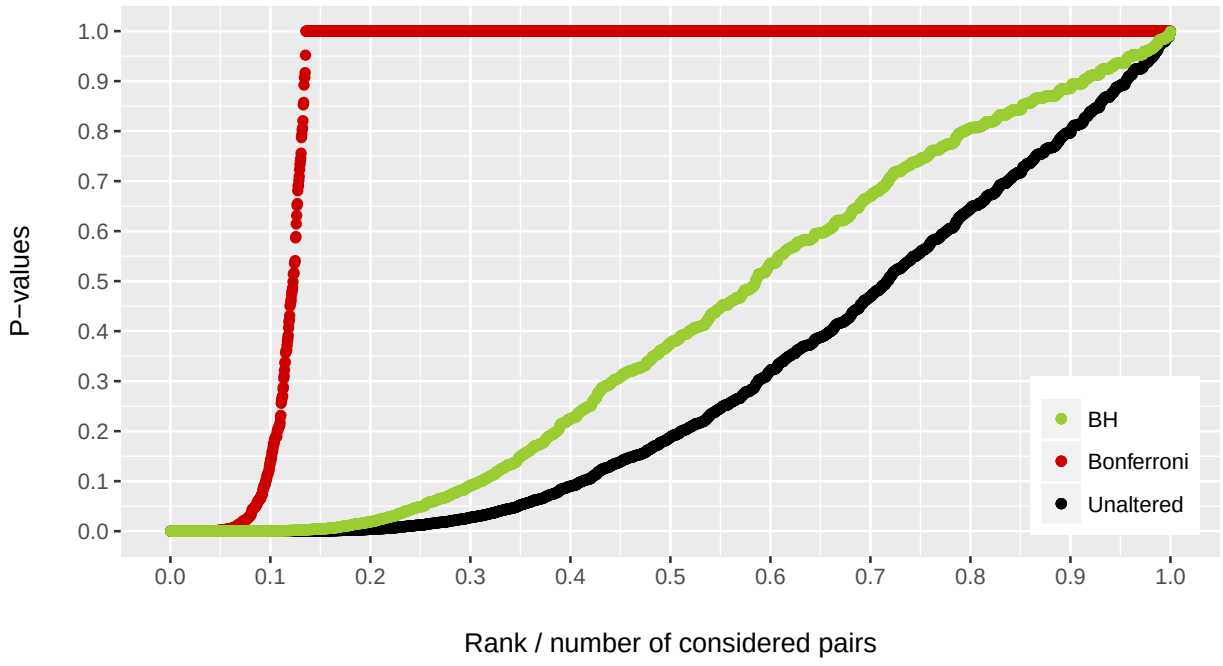


Figure 2.4: P-values distribution according to the relative rank of comparison pairs.
"Rank" stands for the ascending order of the sorted r_s P-values between the standardized ARd and VFd, gathered from all possible combinations between AR and VF protein sub-families depicted in the databases. The "number of considered pairs" designates that, from all possible associations between AR and VF protein sub-families (3813), the r_s P-value could only be calculated for 2716 association pairs, being these the considered ones. As one can see, the r_s P-values when corrected by the Benjamini-Hochberg procedure ("BH", green line), and plotted against the relative rank of the association pairs, slightly deviate from the curve portrayed by the original r_s P-values ("Unaltered", black line). Whereas the Bonferroni correction ("Bonferroni", red line) appears to be too stringent, abruptly reducing the number of significant r_s P-values.

The correlation coefficient (r) and the slope of the linear fits were also calculated as to access the strength and direction of the linear relationship between the standardized ARd and VFd on the generated scatterplots. Associations with $r_s \geqslant 0.5$, that also had $r \geqslant 0.5$ and a r_s P-value < 0.001 were considered valid upon further visual inspection of the points' (metagenomes) distribution. Welch Two Sample *t*-tests [201] were also performed on the standardized ARd/VFd ratios for each metagenome plotted against the age of the human host – depicted in the cohort enclosing human gut metagenomes – as to to test if the given populations have equal means concerning the latter ratios. The former test is but a derivation of Student's *t*-test, being acknowledged as more reliable when the two tested samples have unequal sizes [202]. All statistical analysis was conducted with the R programming language (version 3.2.2). Plots were generated using R's *ggplot2*, *grid*, *gtable* and *scales* packages.

3. Results

3.1 Antibiotic resistance (AR) protein families in the metagenomes

As a first step, the present work asked the two following questions: (i) for a given number of cluster representative sequences of a metagenome (its "size"), how much does the diversity number of antibiotic resistance protein families' homologues (that is, the number of ARd) vary?; and (ii) to what extent does ARd increase when the size of the metagenomes also increases (Equation 2.2)?

In order to answer these questions, we used a dataset of environmental metagenomes issuing from diverse ecosystems and biomes [151] (see Methods). In this dataset, a broad variation in ARd can be seen (Figure 3.1.A), even for a given fixed metagenome size. For example, when the size of a metagenome is about 1.75e+05, the number of ARd can vary between almost zero (chicken cecum and cow rumen) to about 3000 (acid mines drainage biofilms). However, there is a close to linear relationship between the number of ARd and the size of metagenomes. Indeed, upon drawing a regression line fitted through zero, representing the linear relationship between the two variables, one retrieves $r > 0.75$, $r_s > 0.64$, r_s P-value < 0.001, and slope = $\alpha \approx 0.0048$. A regression line for the subset of metagenomes belonging to the human gut biome has also been drawn.

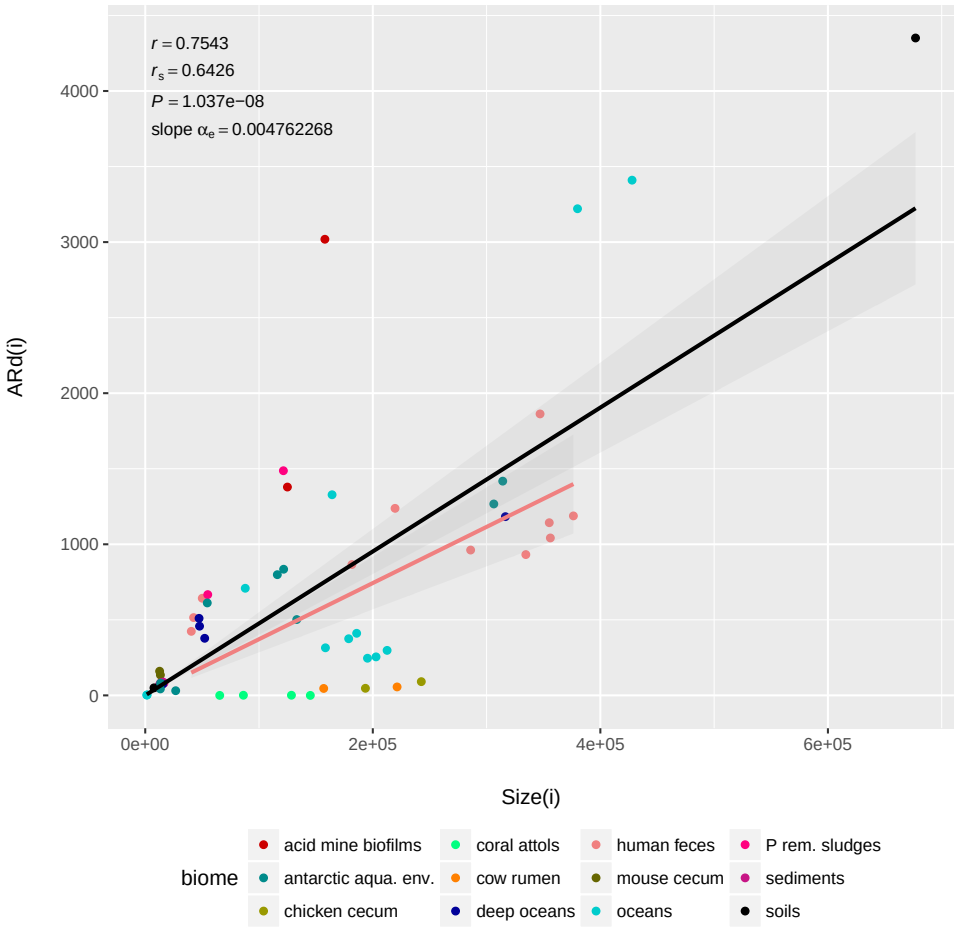
Another dataset composed solely of human gut metagenomes from 110 healthy individuals with ages ranging from 0.05 to 53 years of age, spanning from different regions of the world such as: USA, Malawi, and Venezuelan Amazon [167], was also studied. In this dataset (Figure 3.1.C) the frequency of ARd is strongly dependent on the metagenomes' size ($r > 0.91$, $r_s > 0.94$, r_s P-value < 0.001).

3.2 Virulence factor (VF) protein families in the metagenomes

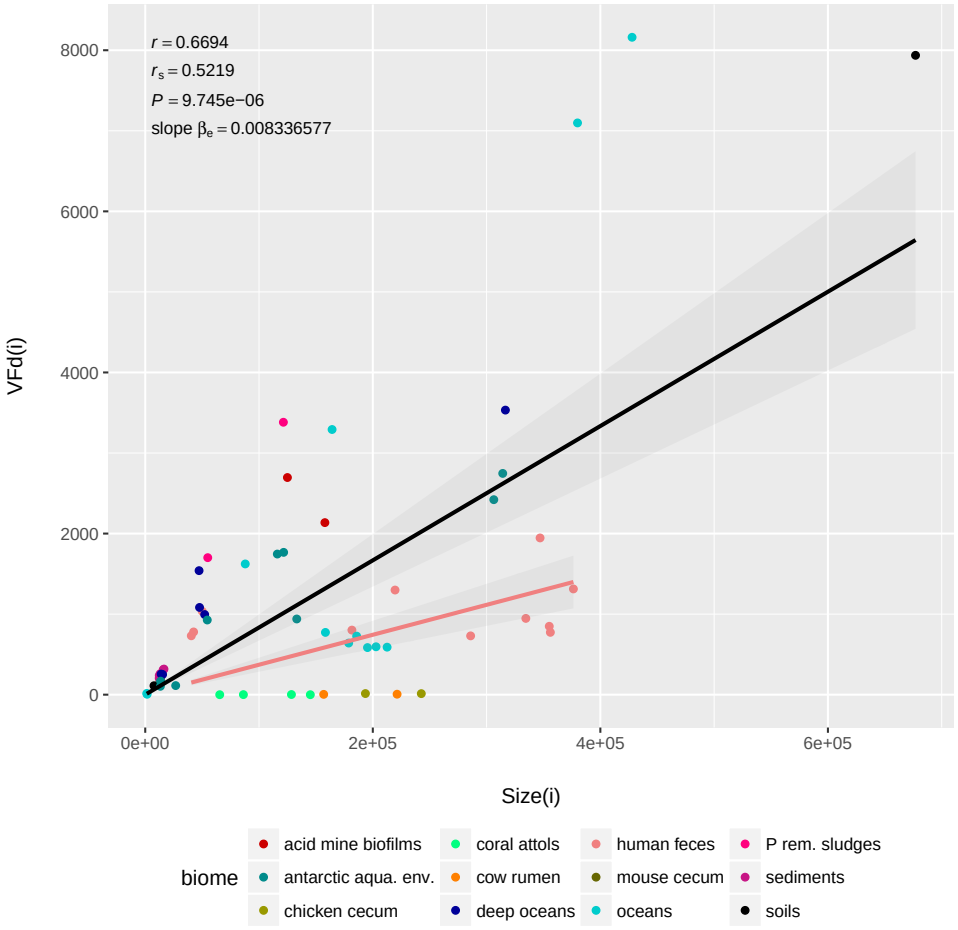
Next, we ought to ask the same questions as those inquired above, but minding virulence factors homologues instead, that is: (i) for a given metagenome size, how much does the number of VFd vary?; and (ii) how much does VFd increase when the metagenomes' size does the same (Equation 2.3)? As to provide answers for these questions, we made use of the same datasets as before.

The metagenomes pertaining to the environmental dataset reveal a great diversity of VFd densities (Figure 3.1.B). For example, when the metagenome size is about 3.7e+05, the number of VFd varies from about 1000 (human faeces) to more than 7000 (ocean). Since virulence may also be associated with the colonization of different types of biomes, besides the context of infection, one can expect different types of these determinants in the environmental microbiomes. In fact, when drawing a regression line fitted through zero one can see $r > 0.66$, $r_s > 0.52$, and r_s P-value < 0.001. However, in metagenomes pertaining to the human gut dataset, a strong correlation between VFd and the size of the metagenomes (Figure 3.1.D) can be observed as well ($r > 0.63$, $r_s > 0.76$, r_s P-value < 0.001), albeit with a bigger r_s (human gut dataset: $r_s = 0.7626$, versus environmental dataset: $r_s = 0.5219$), and a much lower r_s P-value (human gut dataset: r_s P-value = 2.2e-16, *versus* environmental dataset: r_s P-value = 9.745e-06). In the human gut dataset and for VFd, the linear associations portrayed by Venezuela and Malawi seem to be the ones driving the global statistics of r and r_s upwards, since the metagenomes belonging to the USA display a weaker linear association.

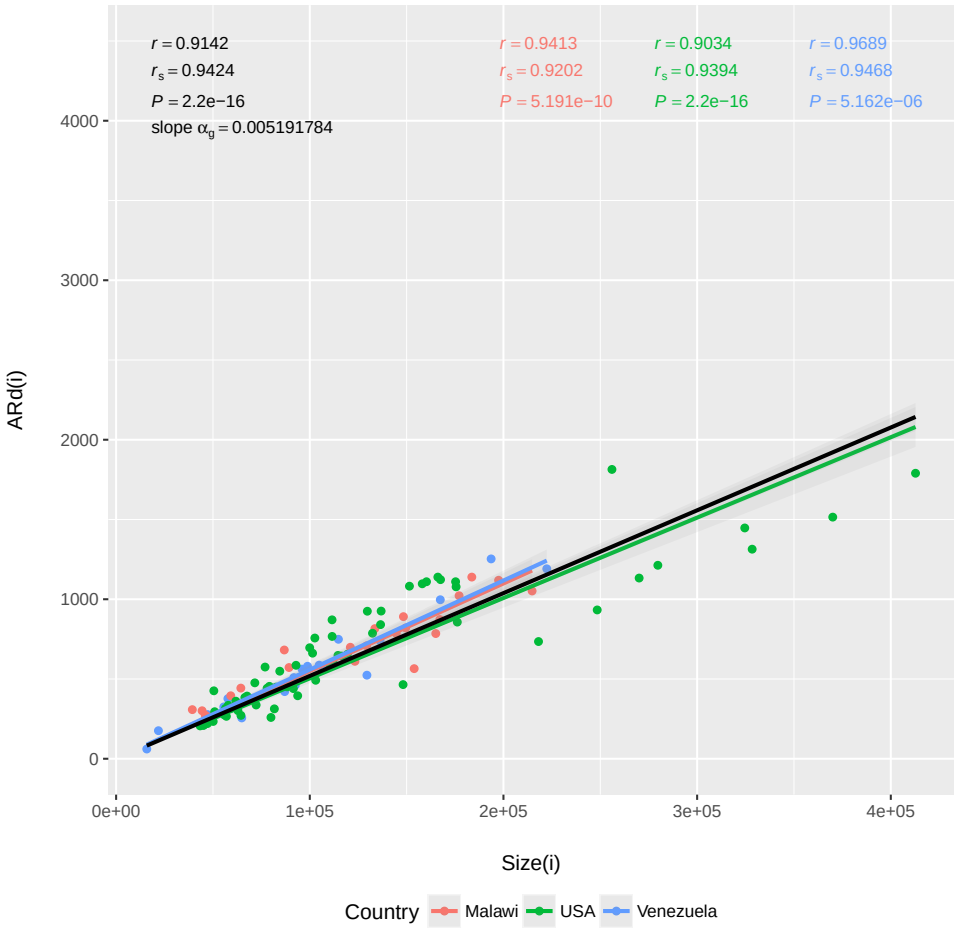
3.1.A


r = 0.7543
r_s = 0.6426
P = 1.037e−08
slope α_e = 0.004762268
ARd(i)
Size(i)
0
1000
2000
3000
4000
0e+00
2e+05
4e+05
6e+05
biome
acid mine biofilms
antarctic aqua. env.
chicken cecum
coral attols
cow rumen
deep oceans
human feces
mouse cecum
oceans
P rem. sludges
sediments
soils

3.1.B


r = 0.6694
r_s = 0.5219
P = 9.745e−06
slope β_e = 0.008336577
VFd(i)
Size(i)
0
2000
4000
6000
8000
0e+00
2e+05
4e+05
6e+05
biome
acid mine biofilms
antarctic aqua. env.
chicken cecum
coral attols
cow rumen
deep oceans
human feces
mouse cecum
oceans
P rem. sludges
sediments
soils

3.1.C


r = 0.9142
r_s = 0.9424
P = 2.2e−16
slope α_g = 0.005191784
r = 0.9413
r_s = 0.9202
P = 5.191e−10
r = 0.9034
r_s = 0.9394
P = 2.2e−16
r = 0.9689
r_s = 0.9468
P = 5.162e−06
ARd(i)
0
1000
2000
3000
4000
0e+00
1e+05
2e+05
3e+05
4e+05
Size(i)
Country
Malawi
USA
Venezuela

3.1.D

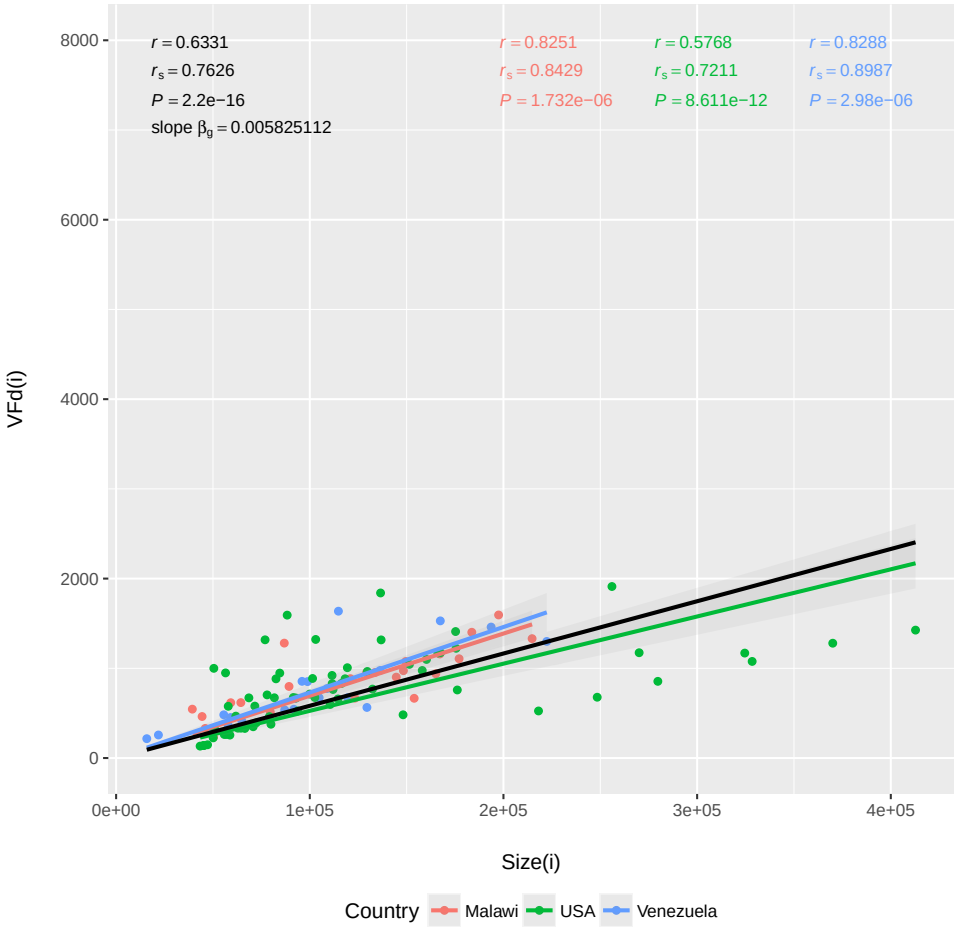

r = 0.6331
r_s = 0.7626
P = 2.2e−16
slope β_g = 0.005825112
r = 0.8251
r_s = 0.8429
P = 1.732e−06
r = 0.5768
r_s = 0.7211
P = 8.611e−12
r = 0.8288
r_s = 0.8987
P = 2.98e−06
VFd(i)
0
2000
4000
6000
8000
0e+00
1e+05
2e+05
3e+05
4e+05
Size(i)
Country
Malawi
USA
Venezuela

Figure 3.1: Distribution of the diversity number of AR and VF protein families by metagenome.
The vertical axes represent the total diversity count of AR and VF protein families' homologues (ARd and VFd respectively) present in the metagenomes. The horizontal axes represent the "size" of the metagenome, that is, the number of cluster representative sequences - see Methods. A and B each represent the 64 environmental metagenomes. The black line represents the simple linear regression of best fit for all the environmental metagenomes, where the grey shading is the 95% confidence interval. The points are scattered showing that the diversity of AR and VF protein families can vary greatly from metagenome to metagenome. The pink line represents the linear regression of best fit for the human faeces metagenomes subset. In C and D each dot refers to one of 110 metagenomes pertaining to the human gut dataset. The red (Malawi), green (USA) and blue (Venezuela) lines depict the linear regressions, respectively, and the black line depicts the regression line for all metagenomes. In this dataset it is evident that the human gut microbiome shares a less diversified set of AR and VF protein families than the environmental metagenomes. Data on the simple correlation coefficient (r), Spearman's rank correlation coefficient (r_s) and the P-value obtained from the latter is shown on all plots. In C and D, statistics pertaining to each individual country are also shown with the colours attributed to latter, accordingly. α_e and β_e are the slopes for the environmental metagenomes (in A and B), whereas α_g and β_g are the slopes for the human gut metagenomes (C and D).

3.3 The AR / VF correlations

The main purpose of the present work is to evaluate the relationship, if any, between ARd and VFd (see Figure 2.3). In order to achieve this goal, we excluded from our analysis all the gene products that were both homologues to AR and VF determinants. Thus avoiding the introduction of a potential bias in the correlation analysis, issuing from all proteins that could both act as an AR determinant and a VF (see Methods).

As to address this relationship, we calculated:

$$\frac{ARd(i)}{\alpha.Size(i)} \qquad \text{(Equation 3.1)}$$

and:

$$\frac{VFd(i)}{\beta.Size(i)} \qquad \text{(Equation 3.2)}$$

for all metagenomes, accordingly depicting the standardized counts for ARd and VFd (see Methods). In Figure 3.2 one can see that the two variables are positively correlated: an environmental metagenome with more (less) VFd than expected also has more (less) ARd than expected. Although there is a large-scale of ARd ($r > 0.75$, $r_s > 0.64$, r_s P-value < 0.001) (Figure 3.1.A) and of VFd ($r > 0.66$, $r_s > 0.52$, r_s P-value < 0.001) (Figure 3.1.B) in these metagenomes, there is a strong correlation between ARd and VFd ($r > 0.75$, $r_s > 0.82$, r_s P-value < 0.001)(Figure 3.2). Despite the fact that the values for ARd and VFd are lower in the metagenomes belonging to the human faeces biome (see Figure 3.1.A and 3.1.B), one can witness a steeper linear fit slope belonging to metagenomes that pertain to the human faeces biome, relatively to the one portrayed by all environmental metagenomes.

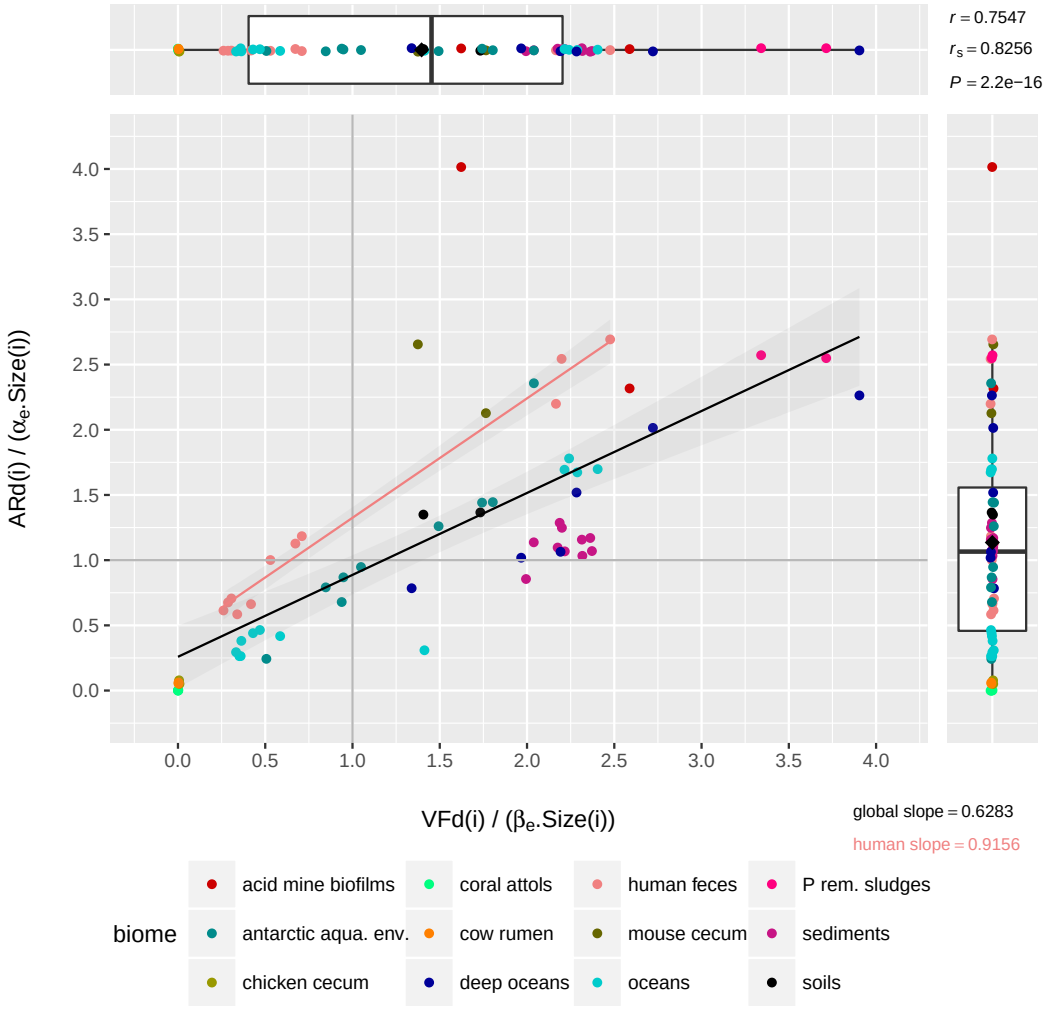
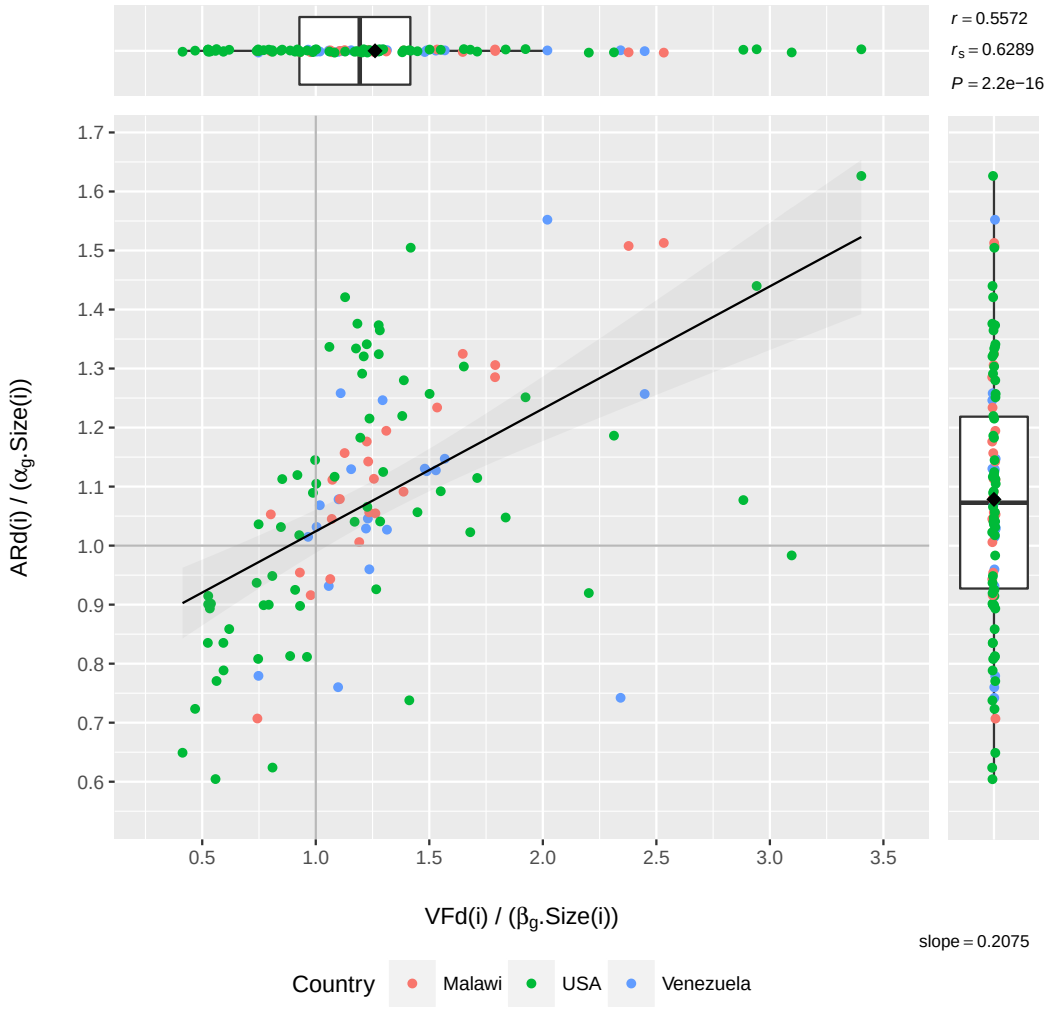


Figure 3.2: Distribution of AR by VF protein diversity in environmental metagenomes.
Scatter plots with marginal boxplots of Equation 3.1 *versus* Equation 3.2 of each metagenome, where α_e and β_e are the slopes for the environmental metagenomes, calculated in Figures 3.1.A and 3.1.B, respectively. The black line represents the simple linear regression of best fit, where the grey shading is the 95% confidence interval. The black line in the box plot represents the median, and the black diamonds represent the mean. The correlations seem to depend on the biome. The pink line represents the linear regression of the human faeces metagenomes subset, which is steeper than the one depicted by the totality of environmental metagenomes. There is a significant correlation involving the 12 different environmental biomes (n = 64, r = 0.7547, r_s = 0.8256, r_s P-value < 0.001, BH post hoc procedure, linear fit slope = 0.6283).

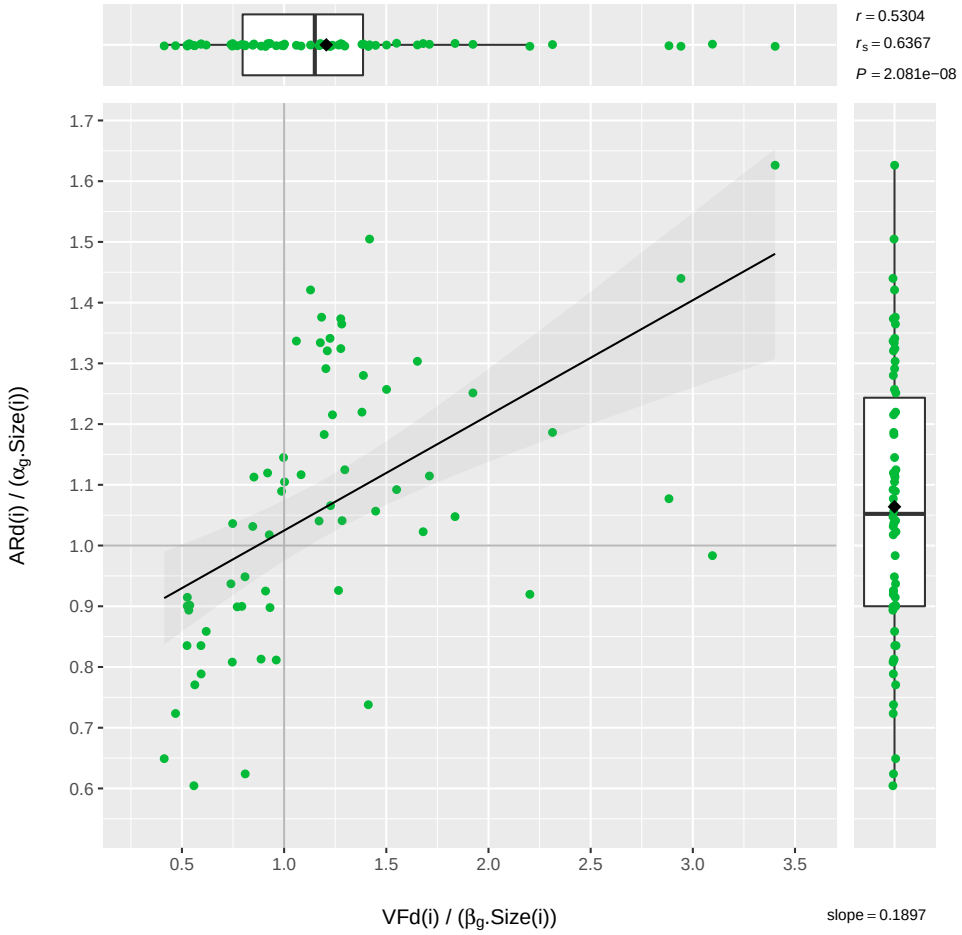
The correlation between ARd and VFd is weaker in the human gut metagenomes dataset ($r > 0.55$, $r_s > 0.62$, r_s P-value < 0.001) (Figure 3.3.A) than in the environmental samples ($r > 0.75$, $r_s > 0.82$, r_s P-value < 0.001) (Figure 3.2). However, we can distinguish different trends when taking the geographical location of the human populations under study into consideration. The biggest contribution to this graph comes from the North American samples, which account for 66/110 (60%) of the individuals (Figure 3.3.B). In order to better understand the results, we can divide the graph into quadrants, using the axes set by coordinates (1,1) (see Methods). One can see that there are many points (metagenomes) located on the first and third quadrant following the regression line. These points represent the metagenomes for which there is a good correlation between the number of ARd and VFd; that is, both ARd and VFd are either in excess or in deficit, for a given size. Nevertheless, some points fall in the second and fourth quadrants. Points that fall in the second quadrant correspond

to metagenomes for which there are more ARd than expected, but a deficit of VFd. The points that fall in the fourth quadrant pertain to metagenomes for which an increment in VFd is followed by a decrease of ARd, not by its increase. When comparing metagenomes collected from Venezuelans (21/110, or 19% of the metagenomes of the dataset) and Malawians (23/110, or 21%), we can see two completely different scenarios (Figures 3.3.C and 3.3.D). In the Amerindian gut metagenomes there is no statistically significant correlation between ARd and VFd ($r > 0.32$, $r_s > 0.45$, r_s P-value = 0.04111). This result presents itself as quite interesting, when we compare it to the Malawian gut metagenomes where there is a very strong correlation between ARd and VFd ($r > 0.90$, $r_s > 0.87$, r_s P-value < 0.001). The slope of the regression line depicting the association between ARd and VFd belonging to Malawian human gut metagenomes (Figure 3.3.D) is about two times steeper than that of the Amerindian (Figure 3.3.C) and USA metagenomes (Figure 3.3.B). That is, in the Malawian population, for a given number of VFd, its respective number of ARd is about twofold that of the other two countries.

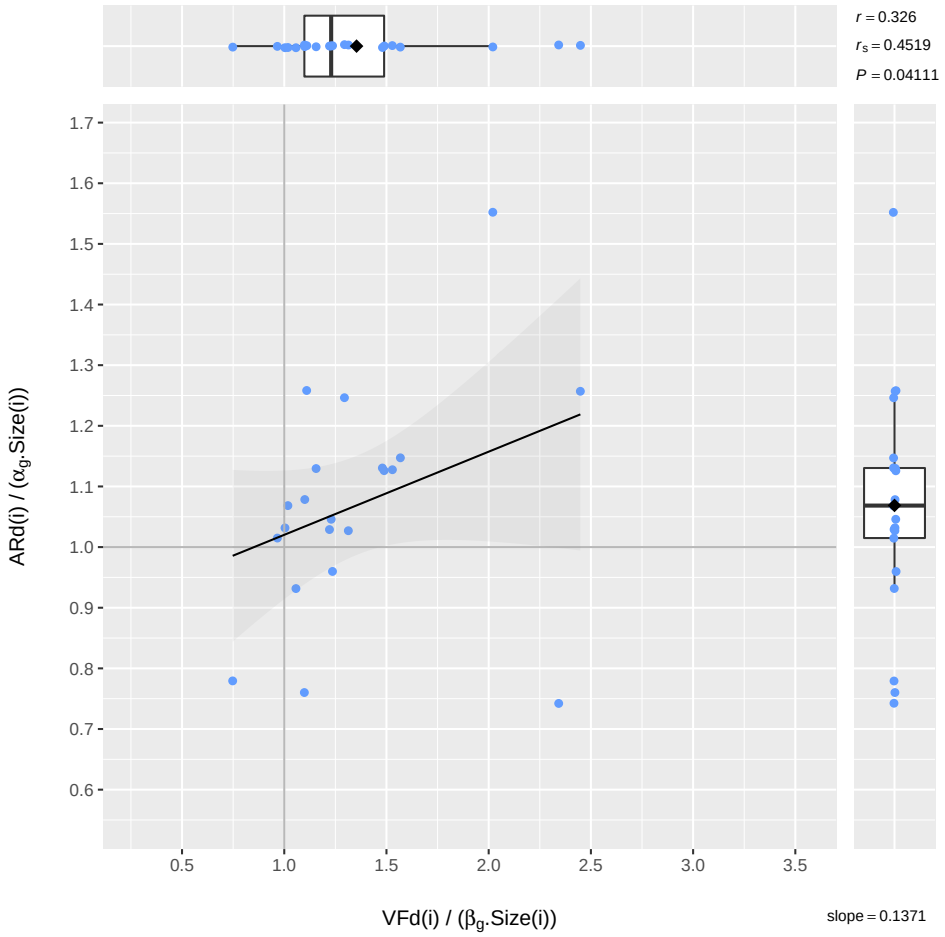
3.3.A



3.3.B


r = 0.5304
rs = 0.6367
P = 2.081e−08
ARd(i) / (αg.Size(i))
VFd(i) / (βg.Size(i))
slope = 0.1897

3.3.C


r = 0.326
rs = 0.4519
P = 0.04111
ARd(i) / (αg.Size(i))
VFd(i) / (βg.Size(i))
slope = 0.1371

3.3.D

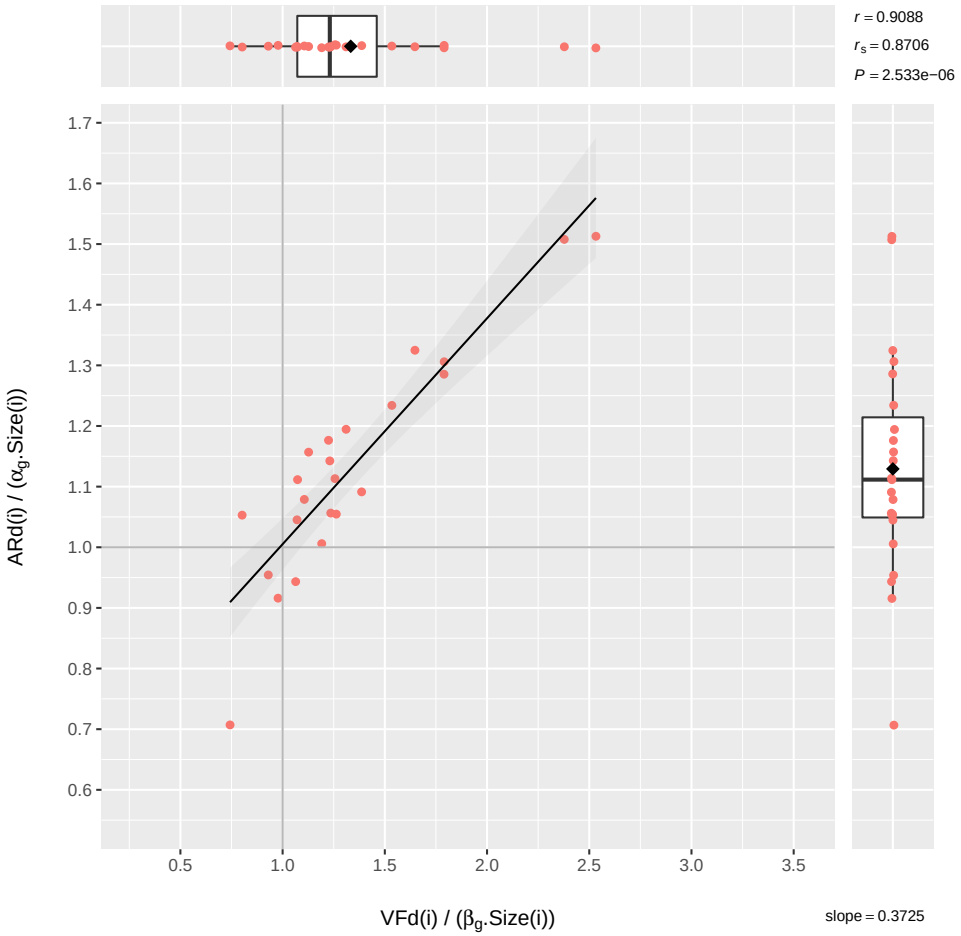


Figure 3.3: Distribution of AR by VF protein diversity in Human gut metagenomes.
Scatter plots with marginal boxplots of Equation 3.1 *versus* Equation 3.2 of each metagenome where α_g and β_g are the slopes calculated in Figures 3.1.C and 3.1.D, respectively. A linear regression line of best fit is represented in each graph, where the grey shading is the 95% confidence interval. The black line in the box plot represents the median and the black diamonds represent the mean. Each colored dot represents a metagenome, and the grey lines parallel to the axes represent the (1,1) reference coordinates. A) Significant correlation involving metagenomes of all sampled human populations – Malawi (red), USA (green), Venezuela (blue), respectively (n = 110, r = 0.5572, r_s = 0.6289, r_s P-value $<$ 0.001, BH post hoc procedure, linear fit slope = 0.2075). B) Significant correlation involving gut metagenomes of USA individuals (n = 66, r = 0.5304, r_s = 0.6367, r_s P-value $<$ 0.001, BH post hoc procedure, linear fit slope = 0.1897). C) Non-significant correlation involving gut metagenomes of Venezuelan individuals (n = 21, r = 0.326, r_s = 0.4519, r_s P-value $>$ 0.001, BH post hoc procedure, linear fit slope = 0.1371). D) Strong significant correlation involving gut metagenomes of Malawian individuals (n = 23, r = 0.9088, r_s = 0.8706, r_s P-value $<$ 0.001, BH post hoc procedure, linear fit slope = 0.3725).

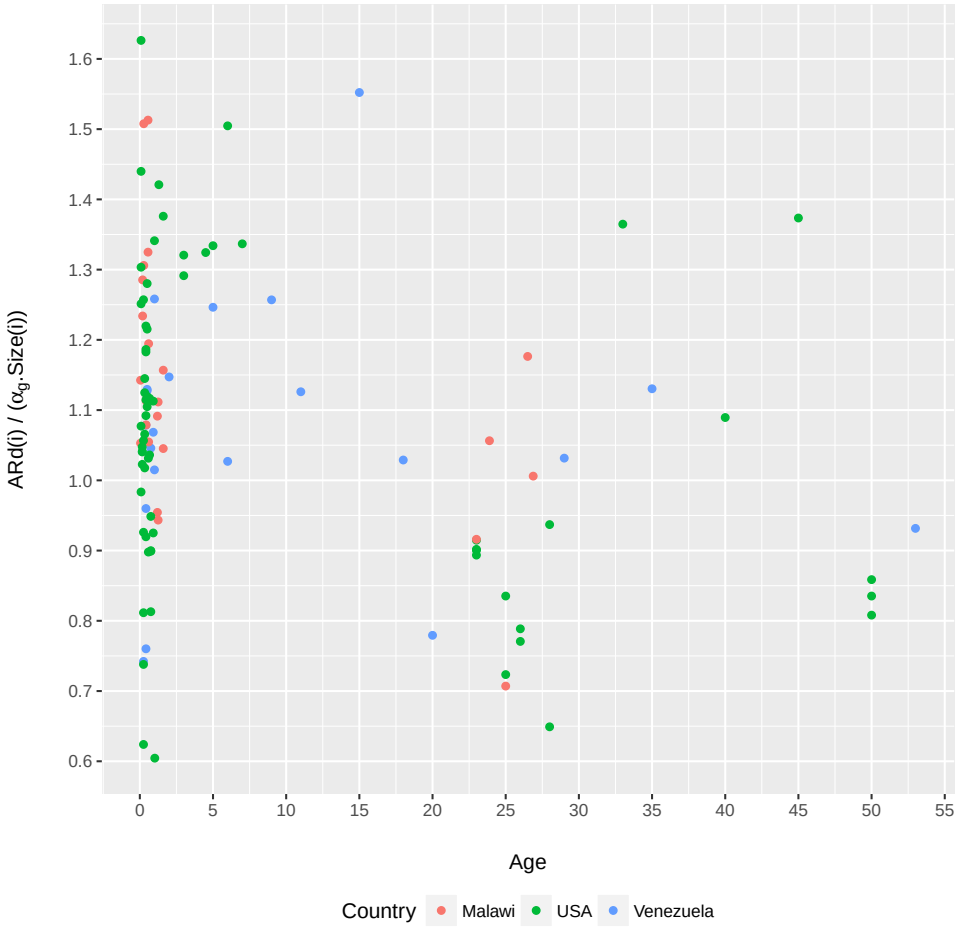
3.4 AR and VF throughout the age of the human gut metagenomes hosts

At this point we wondered whether the associations between AR and VF families, once established in the human gut microbiome, would remain or not, as a consequence of the genetic organization of the mobile genetic elements, such as integrons and transposons. For that, we calculated the standardized ARd/VFd ratios for each metagenome:

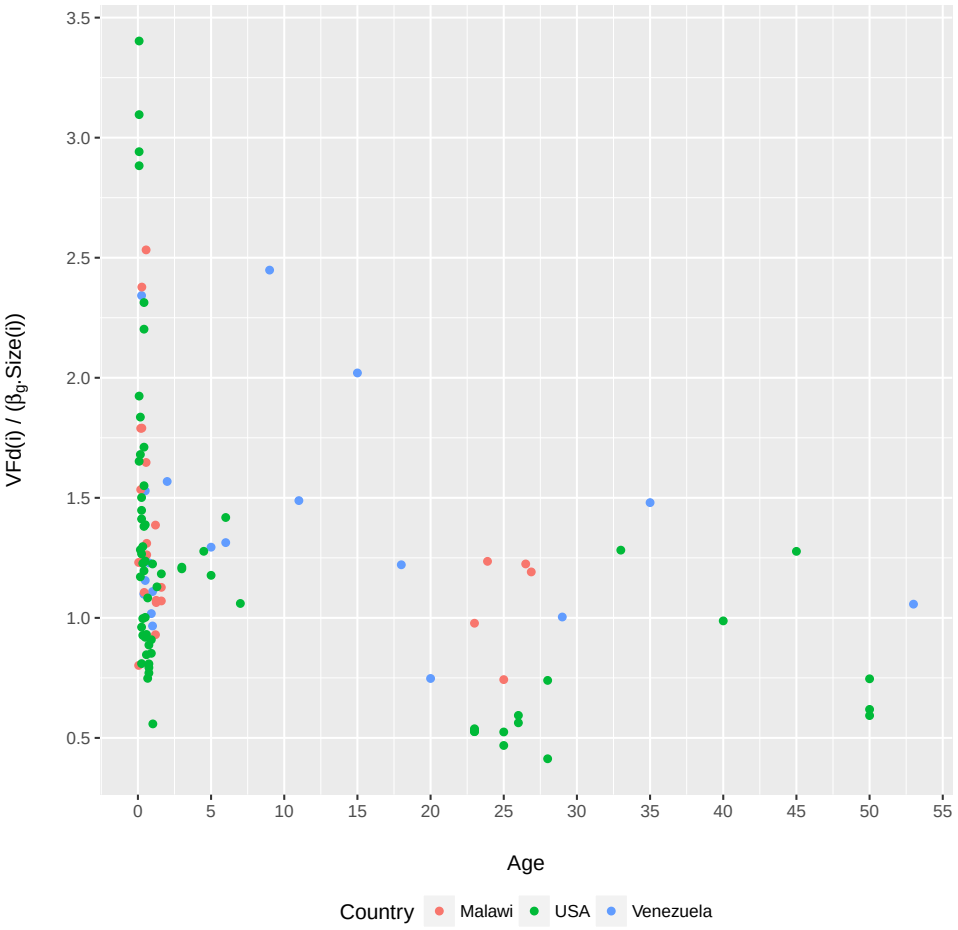
$$\frac{\frac{ARd(i)}{\alpha_g.Size(i)}}{\frac{VFd(i)}{\beta_g.Size(i)}} = \frac{\frac{ARd(i)}{\alpha_g}}{\frac{VFd(i)}{\beta_g}} \qquad \text{(Equation 3.3)}$$

and plotted them against the age of the human host. The gut microbiota is established around the age of three. From this age forward the microbial diversity seems to be stable [167]. Both the ARd and VFd counts seem to decrease with the subjects' age (Figures 3.4.A and 3.4.B). Nevertheless the ARd/VFd ratios seem to remain rather stable throughout the subjects' life in the Amerindian and Malawian gut metagenomes (Figure 3.4.C), despite the decline of individual diversity, both of ARd and VFd with age, and the small number of Malawian individuals sampled after the three-year-old age mark. Furthermore, the ARd/VFd ratio means seem to be quite dissimilar when comparing the USA population against the Malawians (USA ARd/VFd ratio mean = 1.029, Malawi ARd/VFd ratio mean = 0.888, Welch two-sample t-test P-value < 0.01), and the USA population against the Venezuelan Amerindians (USA ARd/VFd ratio mean = 1.029, Venezuela ARd/VFd ratio mean = 0.8399, Welch two-sample t-test P-value < 0.01). Nevertheless, Venezuelans and Malawians possess quite similar ARd/VFd ratio means (Venezuela ARd/VFd ratio mean = 0.8399, Malawi ARd/VFd ratio mean = 0.888, Welch two-sample t-test P-value = 0.3682). It comes as important to notice that although Welch two-sample t-test assumes normality of the population's distribution, it has been reported to remain quite robust even when these distributions present themselves as skewed [203], such as the one currently being presented (Figure 3.4.C). Moreover, upon "magnifying" on the first three years of life (Figure 3.4.D), one can see that the ARd/VFd ratios increase abruptly during the first year of life, and that these ratios reach their peak during this time period, later stabilizing past the age of three, at least for the Venezuelan and Malawian gut metagenomes (Figure 3.4.C).

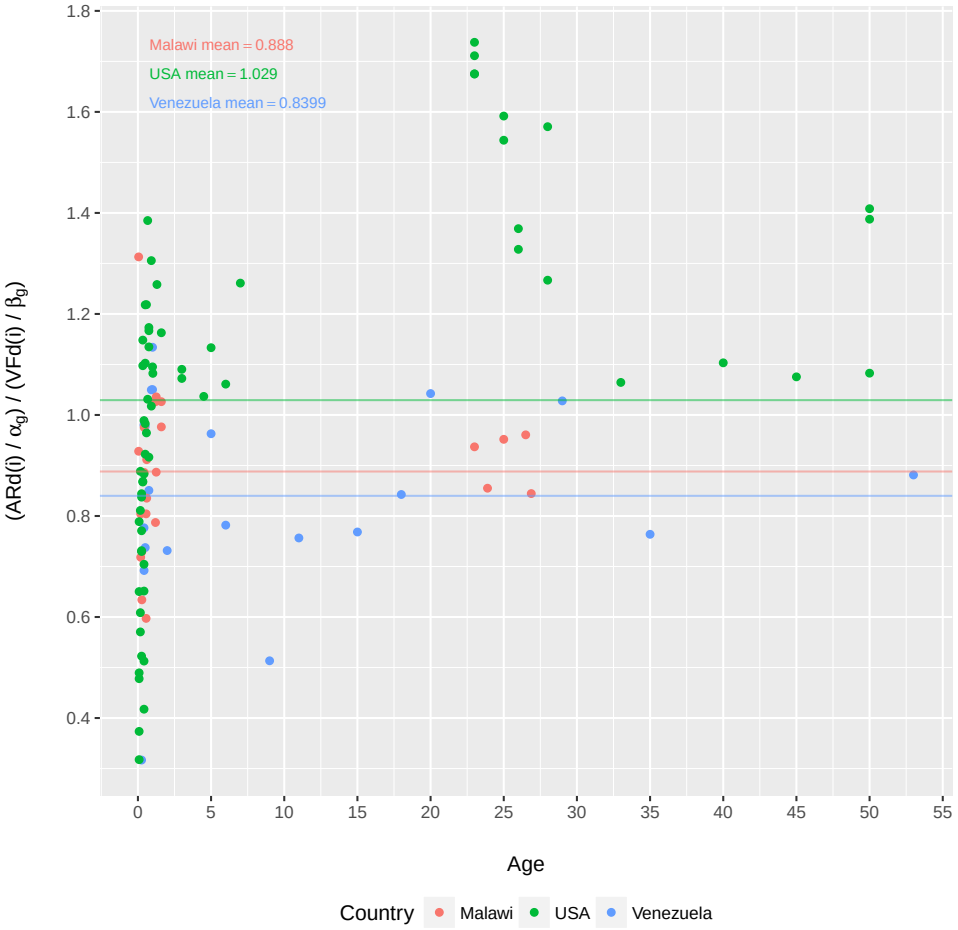
3.4.A



3.4.B


VFd(i) / (βg.Size(i))
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0
5
10
15
20
25
30
35
40
45
50
55
Age
Country
Malawi
USA
Venezuela

3.4.C


Malawi mean = 0.888
USA mean = 1.029
Venezuela mean = 0.8399
(ARd(i) / αg) / (VFd(i) / βg)
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0
5
10
15
20
25
30
35
40
45
50
55
Age
Country
Malawi
USA
Venezuela

3.4.D

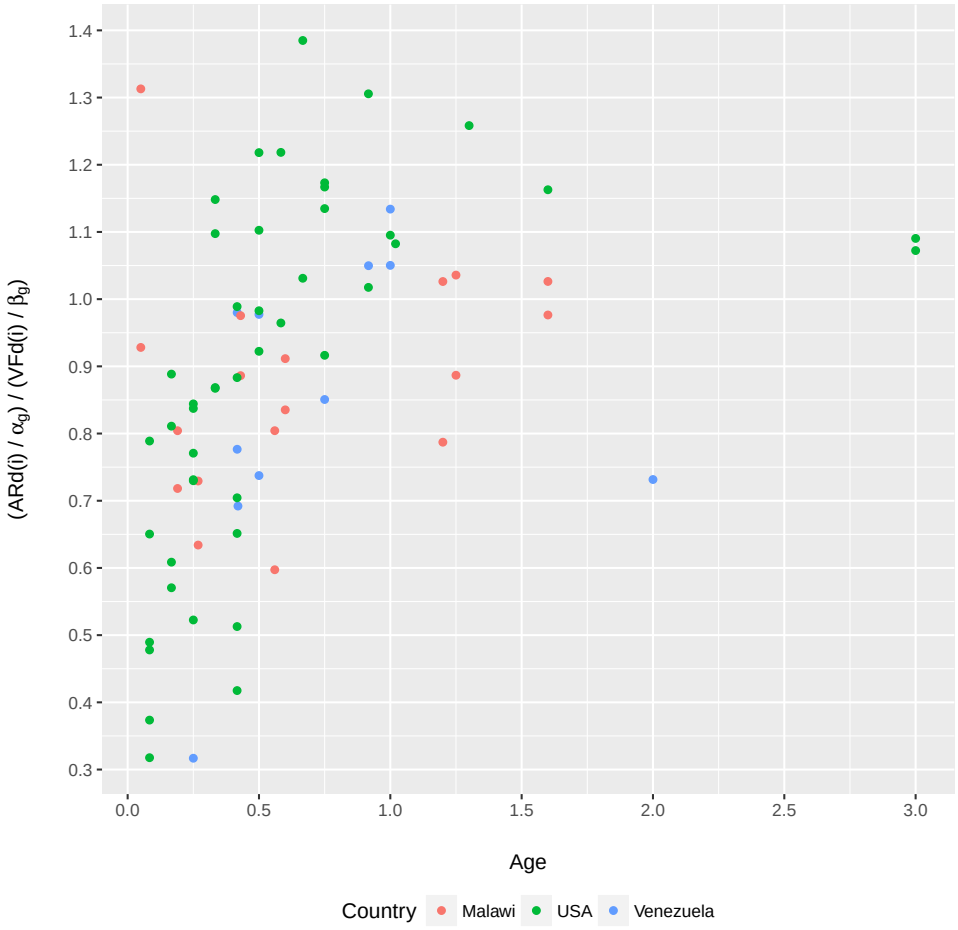


Figure 3.4: ARd/VFd ratios in human gut metagenomes throughout the age of the individuals.
Distribution of Equation 3.1 (A) and Equation 3.2 (B) of each human gut metagenome (where α_g and β_g are the slopes calculated in Figures 3.1.C and 3.1.D, respectively) across the ages (0.05 to 53 years) of 110 individuals from: Malawi (red), USA (green), Venezuela (blue). C) Distribution of the ARd/VFd ratio (Equation 3.3) across the entire individual's age span (0.05 to 53 years); and (D) a zoom to the first three years of life. In Figure C the horizontal lines represent, for each population, the mean of the ratios. This mean is higher for the US individuals (mean = 1.029), and there are some points that represent young adults whose gut microbiome accumulates AR genes over VF ones (i.e.: above the $y = 1$ threshold). The individuals from the USA bear a quite significant difference upon comparison against the other two populations (Welch two sample t-test P-value < 0.01, on both comparisons). On the other extreme, Amerindians from Venezuela show the lowest ratios regardless of age (mean = 0.8399). Malawians show an intermediate mean of 0.888, being similar to that of Amerindians (Welch two sample t-test P-value = 0.3682).

3.5 The co-representation of AR and VF belonging to the cell envelope

Under the same context of antibiotic exposure, one can conceive that co-selection of AR determinants and VFs is presumably taking place. We wondered, however, which were the genetic traits that were more prone to this effect. Thus, after all possible associations between subfamilies of AR and VF protein pairs had been generated (123 * 31 = 3813) for both environmental and human gut metagenomic cohorts, a r_s cut-off of 0.5 was accordingly applied as to filter the best correlations. Upon gathering the latter, it was evident that for both datasets, the vast majority falls into the functional category of multi-drug efflux pumps (AR determinants) associated with either secretion systems, as well with iron uptake and adhesion mechanisms (VFs), respectively (Table 3.1). Hence, one can ascertain that amongst the most representative associations between AR and VF traits are those belonging to the cell

envelope and the general secretion mechanisms.

Table 3.1: Best correlations between AR and VF determinants on the sampled metagenomic datasets.

AR mechanism (FASTA file)	VF mechanism (FASTA file)	r	r_s	r_s P-value (BH corrected)
Environmental:				
ABC Antibiotic Efflux Pump	Heme-mediated Iron Uptake	0.8785	0.8598	2.787245e-14
msbA	Heme-mediated Iron Uptake	0.8602	0.8738	2.787245e-14
macB	Heme-mediated Iron Uptake	0.8415	0.8889	2.787245e-14
macB	Type VI secretion system & effectors	0.8017	0.8367	2.787245e-14
macB	Fibronectin-binding proteins	0.7463	0.7549	2.665298e-11
TetM-TetW-TetO-TetS	Fibronectin-binding proteins	0.7375	0.7620	1.301533e-11
macB	Type III secretion system & effectors	0.7307	0.8205	2.787245e-14
msbA	Fibronectin-binding proteins	0.7304	0.7545	2.726477e-11
msbA	Others	0.7005	0.7216	5.711802e-10
TetM-TetW-TetO-TetS	Others	0.6594	0.7308	2.603417e-10
Human Gut:				
MFS Antibiotic Efflux Pump	Type VI secretion system & effectors	0.8101	0.6806	6.008628e-14
MFS Antibiotic Efflux Pump	Siderophore-mediated Iron Uptake	0.7489	0.7353	4.979333e-14
ABC Antibiotic Efflux Pump	Heme-mediated Iron Uptake	0.6477	0.6307	4.979333e-14
ABC Antibiotic Efflux Pump	Fibronectin-binding proteins	0.639	0.6603	4.979333e-14
MexW-MexI	Type IV Pili	0.6189	0.5899	4.979333e-14
ABC Antibiotic Efflux Pump	Others	0.6012	0.5639	1.810177e-09
msbA	Fibronectin-binding proteins	0.5538	0.5958	4.979333e-14

4. Discussion

In the present dissertation we show that for a given metagenome's size (total number of non-redundant sequences), the diversity number for homologues of antibiotic resistance protein families (ARd), and the diversity number for homologues of virulence factors protein families (VFd) is quite diverse, when taking the biomes from where the sequenced samples were gathered into comparison. We also demonstrate that, in most metagenomes pertaining to the chosen cohorts, when the values for ARd increase, those for VFd also increase.

The variation in diversity of antibiotic resistance (AR) determinants enclosed by environmental metagenomic samples (Figure 3.1.A) may result from the differential microbial community composition of the metagenomes, whose genetic diversity can be grouped according to the adaptation to its respective environment [151]; but also from the fact that selective pressures for the maintenance of antibiotic resistance genes in environmental microbiomes varies widely from environment to environment [204-206].

Human gut microbiomes share a less diversified repertoire of resistance, than their environmental counterparts, as they have already been subjected to adaptation to the environment encompassed by the gastrointestinal tract. Nonetheless, in human gut metagenomes one can see a very strong correlation between the diversity of AR protein homologues and the metagenome size (Figure 3.1.C), independently of the geographical origin of the given human populations. These similar densities of AR genes might suggest that in human gut microbiomes the diversity of the latter is not influenced by the human lifestyle, such as: diet, medical care, access to antibiotics or other cultural habits, as it had been previously described [207], and that the adaptation to the intestinal tract shapes microbial diversity for AR traits as well. The human gut microbial diversity is established until the age of three, without major interpersonal variations on the microbial composition upon geographic distribution [167]. Geography and cultural traditions are, however, responsible for significant differences in the phylogenetic composition of gut microbiomes pertaining to individuals that originate from different countries, when a broader age-span is under consideration. Being the most pronounced divergences those that occur between the USA and the Malawian and Amerindian gut microbiomes [167].

Attending to virulence, one can assert that there appears to be a wide disparity concerning the diversity and density of virulence factors (VFs) in environmental metagenomes, which might pose as evidence of the plasticity displayed by environmental bacteria in order to adapt to different environmental niches (Figure 3.1.B). On the other hand, human gut microbiomes harbour a less diverse VF repertoire, especially in the samples issuing from the USA (Figure 3.1.D), which seem to indicate an evolution towards adaptation to the human gut, or lower contact with pathogens, eventually due to vaccination and sanitation.

One of the most relevant results shown in this dissertation comes from the correlations between ARd and VFd, whether in environmental metagenomes (Figure 3.2), or in the ones pertaining to the human gut cohort (Figures 3.3). ARd and VFd present themselves as strongly correlated in environmental metagenomes, and in one particular subset of the human gut dataset (Figure 3.3.D). The North American intestinal samples (Figure 3.3.B) show a wide variety of associations between ARd and VFd. There is a statistically significant correlation between these genetic traits still allowing some variation, being graphically translated as dispersion. This result in itself reinforces our hypothesis that antibiotic resistance and virulence are in fact co-associated, as an outcome of previous exposure to antibiotics. However, there are some interesting exceptions: there are 5/66 (7.5%) dots in the second quadrant of figure 3.3.B that represent metagenomes where there is an accumulation of ARd by VFd. This outcome

might suggest that the mobilization of mobile genetic elements as gene cassettes is taking place amidst these individuals, eventually leading to possible multiresistence phenomena. Analogously, and still appertaining to the same Figure, there are 3/66 (4.5%) dots that fall directly into the fourth quadrant, indicating an accumulation of VFd over ARd. The USA population, like other industrialized countries, is culturally exposed to antibiotics from health care facilities such as hospitals, early discontinuation of antibiotic therapy, as well as an abuse/misuse of antibiotics in non-clinical settings and processes, such as agriculture and livestock production [94,95]. Furthermore this country (out of the three sampled ones) also has the greatest disparity between social classes, which could in due term be translated to different levels of access to medication and medical treatment.

Among the more similar human gut cohorts – Amerindians from Venezuela and Malawians, the different effect of antibiotic exposure is quite clear when taking their respective ARd *versus* VFd correlations into account (Figures 3.3.C and 3.3.D, respectively). This may also be related to the fact that the former populations share less divergent human gut microbiomes than the North Americans [167], which have been reported to possess a distinctive enterotype [208]. Whereas, there is a non-significant weak correlation between ARd and VFd, amongst the metagenomes spanning from the Venezuelan individuals, there is a very strong correlation amongst the Malawian ones. This result is quite remarkable, since it highlights the effect the exposure to antibacterial drugs exerts on the dissemination of antibiotic resistance, and the co-representation of virulence traits within bacterial communities. Moreover, there are 18/23 (78%) Malawian metagenomes in Figure 3.3.D, that fall into the first quadrant of the plot, indicating that the vast majority of the sampled Malawian individuals have more ARd and more VFd than expected if one considered this increase to be solely reliant on the metagenome's size. The collective bacterial microbiome of uncontacted Amerindians has been regarded as a "frozen" relic of a pre-antibiotic era of the human resistome, despite the fact that it has been known to harbour fully functional antibiotic resistance genes [172], while the Malawian gut microbiome appears to be much more exposed, both to antibiotics, and to colonization by pathogens. There is also two times more ARd per VFd in Malawian human gut metagenomes than in the Amerindian ones. Amerindians have no known access to pharmaceutical-grade drugs, as they usually make use of traditional indigenous medicine as to treat diseases, such as infections. Malawi, on the other hand, is one of the poorest countries in Africa; where the majority of the people live with less than one dollar a day, many people cope with AIDS, and where many children suffer from severe malnutrition [WHO, s.d.]. Bearing in mind the fact that nutrition can play a big role on both human gut microbiome and resistome composition [174,207,209,210], one should also heed that in this country, under UNICEF's authority, a program of Ready-to-Use-Therapeutic-Food (RUTF) has been implemented in order to reduce mortality rates amongst children. However, this RUTF often contains antibiotics such as co-trimoxazole, fact that has led several authors to question whether the success of this therapeutic food is due to re-nutrition or due to the addition of antibiotics to the former [211,212], which could very well be affecting the gut's microbial composition of the subjects undergoing this kind of treatment. Furthermore, it has also been stated that there happens to be widespread resistance to almost all antibiotics being empirically used in Malawi due to the lack of routine microbiologic cultures and sensitivity testing procedures [212], but also due to self-medication.

The ARd and VFd counts seem to decrease when taking the age of the sampled individuals into consideration (Figures 3.4.A and 3.4.B, respectively). These results suggest that the mutualistic flora might become somewhat predominant, and resilient. Nevertheless, the ARd/VFd ratios seem to remain somewhat stable throughout the life of the individuals pertaining to the human gut metagenomic dataset (Figure 3.4.C), evidencing a shared increase in diversity of both AR and VF determinants during the first

year of life (Figure 3.4.D), and then settling in attenuated levels past the age of three in the Venezuelan and Malawian individuals, whereas the North American subjects display higher ratios between the ages of 20 and 30. This is an astonishing result because it conveys to some extent that, not only there is an increased association (co-selection/co-representation) between protein families of VF and AR determinants in the human microbiome, but also that this correlation seems to persist in the bacterial community (AR-VF gene pairs endure once settled), as a purported fingerprint of the co-selection process. Horizontal gene transfer and mobilization could explain these facts through the action of mobile genetic elements such as transposons and integrons. This ratio is much higher in the USA, indicating an association above what's expected (as shown by the lines of the ARd/VFd averages). Despite the fact that there is less phylogenetic diversity within this group, some metagenomes tend to keep high AR / VF gene product rates. The selective pressure by the presence of different antibiotics might be stimulating the mobilization of gene cassettes or transposons (leading to multidrug resistance).

Between all the possible statistical AR and VF gene families' correlations, the best matches are amongst those belonging to the bacterial cell envelope (Table 3.1). This result is not surprising, as many proteins that belong to the secretory system or the secretome (the collective sum of all secreted proteins from a given species) itself, are frequently encoded on mobile genetic elements [128,213]. The most relevant cases of resistance mechanisms to antibiotics are components of multidrug efflux pumps and export machineries; and virulence factors that belong to mechanisms of iron uptake (heme uptake proteins and siderophores), as well as secretion systems involved in invasion and adherence to host cells. There are a few bibliographic records of these combinations. For instance, biofilm production is known to be related with the activity of some VFs, like type IV pili, and at least one reported type IV secretion system [46,136,147]. Furthermore, biofilm formation in a *S. aureus* methicillin resistant (MRSA) strain has been known to be essentially reliant on the activity of fibronectin binding proteins [53]. Multidrug efflux pumps also have direct implications with the formation and maintenance of such matrices [137], as well as promoting indirect interplay through quorum sensing modulation, which in due course is known to control biofilm differentiation and further expression of several VFs [149]. There's also a two-component regulatory system involved in biofilm formation, and regulation of resistance and virulent traits [138]. Additionally, it has also been settled that quinolone resistant *S. aureus* strains up regulate the production of fibronectin binding proteins when subjected to sub-lethal dosages of ciprofloxacin [214], which, even though it might pose as an indirect recount of co-selection, still reminiscences on the plasticity that can be portrayed by VFs when antibiotic compounds are present. It has also been acknowledged that physiological levels of some cations present within the host, that in such a manner end up acting as signalling agents of entrance inside the environmental milieu encompassed by the latter, promote the up regulation of genes encoding putative efflux transporters, oxidoreductases, and mechanisms of iron uptake either in *A. baumannii* [215], as in *Burkholderia cenocepacia* [216], which could explain the co-association of iron acquisition systems with those of multidrug efflux pumps. On the other extreme of the scale, associations between all AR subfamilies and the specific VF subfamilies that encode for toxins, bore no statistical significance.

Despite the fact that correlation does not always indicate causality, all the results gathered by our team, as well as the underlying nature of the associations described thus far, has led us into the conclusion that we are in fact in the presence of a causative relationship between antibiotic resistance and virulence factors, specially in the metagenomes issuing from the human gut cohort.

5. Conclusive Remarks

To our knowledge, the present dissertation poses as the first piece of evidence on the co-representation of AR and VF determinants amongst environmental [151] and human gut [167] metagenomic datasets. As far as we know, it is also the first bibliographical instance where the diversity of both types of determinants has been statistically addressed together using metagenomic data.

The methodology employed herein settles on a scalable, and reusable, workflow that made use of a well-established data-retrieving platform [160]; two publicly available and widely acknowledged protein databases [105,198]; an open source, freely available, and highly renown similarity search algorithm [187]; as well as several data-mining and statistical programming languages customarily used in computational biology research.

Future research venturing into larger metagenomic datasets could probably make use of a faster similarity search algorithm [193] in order to infer homology, while minimizing computational time and requiring less computational power, perhaps relying on even more stringent or additional cutoff criteria, as to benefit from the preceding enhancements without compromising on the reliability and veracity of the resulting alignments.

During this work we came to some conclusions. Firstly, our results confirm that both AR and VF determinants are widely disseminated throughout numerous environments, with special emphasis on the niche encompassed by the human gastrointestinal tract of 110 healthy individuals. Secondly, they also suggest that even though bacterial communities dwelling in the environment have a greater variation of ARd and VFd when taking the metagenome's size into account, the human gut bacterial communities possess a very strong linear dependency when attending to ARd's distribution on the size of the metagenomes, and a strong linear relationship concerning VFd against the metagenome's size. Additionally, we report that the standardized counts for ARd and VFd portray a very strong association in environmental metagenomes, where the metagenomes belonging to the human faeces biome subset display a steeper linear dependency rate than the one portrayed by the totality of this dataset. Moreover, the standardized counts for ARd and VFd are also significantly correlated in the human gut dataset, where the metagenomes pertaining to the individuals spanning from the USA present a wide array of different associations, whereas the Venezuelan samples seem to have no statistically significant association. The Malawian metagenomes depict the strongest correlation and linear relationship out of all three countries, as well as possessing twice as more ARd per VFd than the other two countries represented in the foregoing dataset. Still appertaining to the same metagenomic cohort, we also disclose that both stardardized counts for ARd and VFd seem to decrease with the subject's age, whereas the ARd/VFd ratios appear relatively stable throughout the life of the individuals pertaining to the three sampled countries, evidencing a common precipitous increase in diversity during the first year of life, and then later stabilizing past the age of three, with the exception of a few individuals belonging to the USA. We also present results, which indicate that from the entirety of possible associations between the functionally categorized VF and AR protein sub-families, those that relate to the bacterial cell envelope bore the best statistical correlations.

It should be highlighted that the results presented in this dissertation can only provide evidence for the co-representation of AR and VF determinants amidst environmental and human gut microbiomes sampled worldwide. Since the chosen metagenomic cohorts lack a temporal scale, and were not probed for gene proximity nor inclusion in the same mobile genetic elements, our results cannot attest for the mobilization of both determinants together, i.e.: co-selection and further co-evolution by means of the same genetic vectors. However, the co-representative nature of our results reinforces on the notion of

co-selectivity, as it also lays the path for the establishment of new well-directed experimental inquiries.

Further research is required in order to better understand the nature underlying both ARd and VFd's distribution in human gut metagenomes throughout the host's life, preferably by making use of a metagenomic cohort enclosing several samples of the same individuals during a concise time scale; and also as to confirm if the best correlated AR and VF protein sub-families are being disseminated together in the same mobile genetic elements, thus driving forward co-evolution of both types of determinants.

Our results emphasize on the effects that exposure to antibiotics, emerging either from antimicrobial therapy as from the environment that surrounds us, and food, can exert on the selection for pathogenic bacterial traits amidst the human gut. This occurrence may further drive and shape changes on the gene pool of presumably healthy microbiomes. Such perception comes as very important to all microbiologists, especially those whose research relates to microbial ecology, but also to the entire community of medical and health professionals, as well as those working in the food industry, along with all of us as informed and aware citizens.

References

[1] Woese CR. On the evolution of cells. Proc Natl Acad Sci U S A. 2002 Jun 25;99(13):8742-7.

[2] Allwood AC, Walter MR, Kamber BS, Marshall CP, Burch IW. Stromatolite reef from the Early Archaean era of Australia. Nature. 2006 Jun 8;441(7094):714-8.

[3] N. J. Butterfield, A. H. Knoll, K. Swett. A bangiophyte red alga from the Proterozoic of arctic Canada. Science. 1990 October 5; 250: 104–107.

[4] Ley RE, Peterson DA, Gordon JI. Ecological and evolutionary forces shaping microbial diversity in the human intestine. Cell. 2006 Feb 24;124(4):837-48.

[5] Bäckhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI. Host-bacterial mutualism in the human intestine. Science. 2005 Mar 25;307(5717):1915-20.

[6] Rawls JF, Samuel BS, Gordon JI. Gnotobiotic zebrafish reveal evolutionarily conserved responses to the gut microbiota. Proc Natl Acad Sci U S A. 2004 Mar 30;101(13):4596-601.

[7] Lederberg, J; McCray, AT. 'Ome Sweet 'Omics—a genealogical treasury of words. Scientist 2001;15: 8.

[8] The NIH HMP Working Group, Peterson J, Garges S, et al. The NIH Human Microbiome Project. Genome Research. 2009;19(12):2317-2323.

[9] Jack A Gilbert, Janet K Jansson, Rob Knight. The Earth Microbiome project: successes and aspirations. BMC Biol. 2014; 12: 69.

[10] Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol. 1998 Oct;5(10):R245-9.

[11] Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. Proc Natl Acad Sci U S A. 1998 Jun 9;95(12):6578-83.

[12] Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, et al. Diversity of the human intestinal microbial flora. Science. 2005 Jun 10;308(5728):1635-8.

[13] Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, et al. Enterotypes of the human gut microbiome. Nature. 2011 May 12;473(7346):174-80.

[14] Fargione J, Brown CS, Tilman D. Community assembly and invasion: an experimental test of neutral versus niche processes. Proc Natl Acad Sci U S A. 2003 Jul 22;100(15):8916-20.

[15] Steinhoff U. Who controls the crowd? New findings and old questions about the intestinal microflora. Immunol Lett. 2005 Jun 15;99(1):12-6.

[16] Shreiner, AB, Kao J. Y, Young, V. B. The gut microbiome in health and in disease. Curr Opin Gastroenterol. 2015 January; 31(1): 69–75.

[17] LeBlanc JG, Milani C, de Giori GS, Sesma F, van Sinderen D, et al. Bacteria as vitamin suppliers to their host: a gut microbiota perspective. Curr Opin Biotechnol. 2013 Apr;24(2):160-8.

[18] Flint HJ, Scott KP, Louis P, Duncan SH. The role of the gut microbiota in nutrition and health. Nat Rev Gastroenterol Hepatol. 2012 Oct;9(10):577-89.

[19] Weyrich, L. S. Evolution of the Human Microbiome and Impacts on Human Health, Infectious Disease, and Hominid Evolution, In Reticulate Evolution: Symbiogenesis, Lateral Gene Transfer, Hybridization and Infectious Heredity. Interdisciplinary Evolution Research 3. Gontier, Nathalie and Pombo, Olga. Springer International Publishing Switzerland, 2015. 231-53. Print.

[20] Casadevall A, Pirofski LA. Host-pathogen interactions: redefining the basic concepts of virulence and pathogenicity. Infect Immun. 1999 Aug;67(8):3703-13.

[21] Tamboli CP, Neut C, Desreumaux P, Colombel JF. Dysbiosis in inflammatory bowel disease. Gut. 2004;53(1):1-4.

[22] Chang C, Lin H. Dysbiosis in gastrointestinal disorders. Best Pract Res Clin Gastroenterol. 2016 Feb;30(1):3-15.

[23] Fenner, F. The effects of changing social organization on the infectious diseases of man, p. 3–73. In S. V. Boyden (ed.), The Impact of Civilisation on the Biology of Man. University of Toronto Press, Toronto; 1970.

[24] Haensch S, Bianucci R, Signoli M, Rajerison M, Schultz M, et al. Distinct clones of Yersinia pestis caused the black death. PLoS Pathog. 2010 Oct 7;6(10):e1001134.

[25] Moore S, Thomson N, Mutreja A, Piarroux R. Widespread epidemic cholera caused by a restricted subset of Vibrio cholerae clones. Clin Microbiol Infect. 2014 May;20(5):373-9.

[26] Harris JB, LaRocque RC, Qadri F, Ryan ET, Calderwood SB. Cholera. Lancet. 2012 Jun 30;379(9835):2466-76.

[27] Karlsson EK, Kwiatkowski DP, Sabeti PC. Natural selection and infectious disease in human populations. Nat Rev Genet. 2014 Jun;15(6):379-93.

[28] Masri L, Branca A, Sheppard AE, Papkou A, Laehnemann D, et al. Host-Pathogen Coevolution: The Selective Advantage of Bacillus thuringiensis Virulence and Its Cry Toxin Genes. PLoS Biol. 2015 Jun;13(6):e1002169.

[29] Antonovics J, Boots M, Ebert D, Koskella B, Poss M, et al. The origin of specificity by means of natural selection: evolved and nonhost resistance in host-pathogen interactions. Evolution. 2013 Jan;67(1):1-9.

[30] Brunham RC, Plummer FA, Stephens RS. Bacterial antigenic variation, host immune response, and pathogen-host coevolution. Infect Immun. 1993 Jun;61(6):2273-6.

[31] Webb SA, Kahler CM. Bench-to-bedside review: Bacterial virulence and subversion of host defences. Crit Care. 2008;12(6):234.

[32] Falkow S: The evolution of pathogenicity in Escherichia, Shigella, and Salmonella. In Escherichia coli and Salmonella – Cellular and molecular biology. Edited by Neidhardt FC. Washington, DC: ASM Press; 1996.

[33] Smith H. Biochemical challenge of microbial pathogenicity. Bacteriology Reviews 1968; 32:164–84.

[34] Hoffmann JA. The immune response of Drosophila. Nature 2003, 426:33-38.

[35] Gama JA, Abby SS, Vieira-Silva S, Dionisio F, Rocha EPC. Immune Subversion and Quorum-Sensing Shape the Variation in Infectious Dose among Bacterial Pathogens. Wessels MR, ed. PLoS Pathogens. 2012;8(2):e1002503.

[36] Falkow S. Molecular Koch's postulates applied to microbial pathogenicity. Rev Infect Dis. 1988 Jul-Aug;10 Suppl 2:S274-6.

[37] Falkow S: Molecular Koch's postulates applied to bacterial pathogenicity – a personal recollection 15 years later. Nat Rev Microbiol 2004, 2:67-72.

[38] Hacker J, Blum-Oehler G, Muhldorfer I, Tschape H: Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Mol Microbiol 1997, 23:1089-1097.

[39] Dionisio F, Matic I, Radman M, Rodrigues OR, Taddei F. Plasmids spread very fast in heterogeneous bacterial communities. Genetics. 2002 Dec;162(4):1525-32.

[40] Mahan MJ, Heithoff DM, Sinsheimer RL, Low DA: Assesment of bacterial pathogenesis by analysis of gene expression in the host. Annu Rev Genet 2000, 34:139-164.

[41] Hornef MW, Wick MJ, Rhen M, Normark S: Bacterial strategies for overcoming host innate and adaptive immune responses. Nat Immunol 2002, 3:1033-1040.

[42] Finlay BB, Falkow S. Common themes in microbial pathogenicity revisited. Microbiol Mol Biol Rev. 1997 Jun;61(2):136-69.

[43] Sottile FD, Marrie TJ, Prough DS, Hobgood CD, Gower DJ, et al. Nosocomial pulmonary infection: possible etiologic significance of bacterial adhesion to endotracheal tubes. Crit Care Med. 1986 Apr;14(4):265-70.

[44] Klemm P, Schembri MA. Bacterial adhesins: function and structure. Int J Med Microbiol. 2000 Mar;290(1):27-35.

[45] De Graaf, F. K., Gaastra, W.: Fimbriae of enterotoxigenic Escherichia coli. In: Fimbriae, adhesion, genetics, biogenesis, and vaccines (P. Klemm, ed.), pp.53-83. CRC Press, Boca Raton, 1994.

[46] Craig L, Pique ME, Tainer JA. Type IV pilus structure and bacterial pathogenicity. Nat Rev Microbiol. 2004 May;2(5):363-78.

[47] Källström H, Liszewski MK, Atkinson JP, Jonsson AB. Membrane cofactor protein (MCP or CD46) is a cellular pilus receptor for pathogenic Neisseria. Mol Microbiol. 1997 Aug;25(4):639-47.

[48] Patti, J.M.; Allen, B.L.; McGavin, M.J.; Hook, M. Mscramm-mediated adherence of microorganisms to host tissues. Annu. Rev. Microbiol. 1994, 48, 585–617.

[49] Pankov R, Yamada KM. Fibronectin at a glance. J Cell Sci. 2002 Oct 15;115(Pt 20):3861-3.

[50] Henderson B, Nair S, Pallas J, Williams MA. Fibronectin: a multidomain host adhesin targeted by bacterial fibronectin-binding proteins. FEMS Microbiol Rev. 2011 Jan;35(1):147-200.

[51] Williams RJ, Henderson B, Nair SP. Staphylococcus aureus fibronectin binding proteins A and B possess a second fibronectin binding region that may have biological relevance to bone tissues. Calcif Tissue Int. 2002. 70 (5): 416–21.

[52] Sinha, B.; François, P.P.; Nüße, O.; Foti, M.; Hartford, O.M.; Vaudaux, P.; Foster, T.J.; Lew, D.P.; Herrmann, M.; Krause, K.H. Fibronectin-binding protein acts as Staphylococcus aureus invasin via fibronectin bridging to integrin $\alpha 5 \beta 1$. Cell. Microbiol. 1999, 1, 101–117.

[53] McCourt, J.; O'Halloran, D.P.; McCarthy, H.; O'Gara, J.P.; Geoghegan, J.A. Fibronectin-binding proteins are required for biofilm formation by community-associated methacillin-resistant Staphylococcus aureus strain LAC. FEMS Microbiol. Lett. 2014, 353, 157–164.

[54] Hornef MW, Wick MJ, Rhen M, Normark S: Bacterial strategies for overcoming host innate and adaptive immune responses. Nat Immunol 2002, 3:1033-1040.

[55] High N, Mounier J, Prévost MC, Sansonetti PJ. IpaB of Shigella flexneri causes entry into epithelial cells and escape from the phagocytic vacuole. EMBO J. 1992 May;11(5):1991-9.

[56] Goldberg MB, Theriot JA. Shigella flexneri surface protein IcsA is sufficient to direct actin-based motility. Proc Natl Acad Sci U S A. 1995 Jul 3;92(14):6572-6.

[57] Green ER, Mecsas J. Bacterial Secretion Systems: An Overview. Microbiol Spectr. 2016 Feb;4(1).

[58] Desvaux M, Hébraud M, Talon R, Henderson IR. Secretion and subcellular localizations of bacterial proteins: a semantic awareness issue. Trends Microbiol. 2009 Apr;17(4):139-45.

[59] Abby SS, Cury J, Guglielmini J, Néron B, Touchon M, Rocha EPC. Identification of protein secretion systems in bacterial genomes. Scientific Reports. 2016;6:23080.

[60] Diepold A, Armitage JP. Type III secretion systems: the bacterial flagellum and the injectisome. Philos Trans R Soc Lond B Biol Sci. 2015 Oct 5;370(1679).

[61] Buttner D. Protein export according to schedule: architecture, assembly, and regulation of type III secretion systems from plant- and animal-pathogenic bacteria. Microbiol Mol Biol Rev. 2012 76:262–310.

[62] Young BM, Young GM. YplA is exported by the Ysc, Ysa, and flagellar type III secretion systems of Yersinia enterocolitica. J Bacteriol. 2002 Mar;184(5):1324-34.

[63] Young GM, Schmiel DH, Miller VL. A new pathway for the secretion of virulence factors by bacteria: the flagellar export apparatus functions as a protein-secretion system. Proc Natl Acad Sci U S A. 1999 May 25;96(11):6456-61.

[64] Cianfanelli FR, Monlezun L, Coulthurst SJ. Aim, Load, Fire: The Type VI Secretion System, a Bacterial Nanoweapon. Trends Microbiol. 2016 Jan;24(1):51-62.

[65] Russell AB, Peterson SB, Mougous JD. Type VI secretion system effectors: poisons with a purpose. Nat Rev Microbiol. 2014 Feb;12(2):137-48.

[66] Lemichez E, Barbieri JT. General aspects and recent advances on bacterial protein toxins. Cold Spring Harb Perspect Med. 2013 Feb 1;3(2):a013573.

[67] do Vale A, Cabanes D, Sousa S. Bacterial Toxins as Pathogen Weapons Against Phagocytes. Front Microbiol. 2016;7:42.

[68] Vilches S, Wilhelms M, Yu HB, Leung KY, Tomás JM, et al. Aeromonas hydrophila AH-3 AexT is an ADP-ribosylating toxin secreted through the type III secretion system. Microb Pathog. 2008 Jan;44(1):1-12.

[69] Carbonetti NH. Pertussis toxin and adenylate cyclase toxin: key virulence factors of Bordetella pertussis and cell biology tools. Future Microbiol. 2010 Mar;5(3):455-69.

[70] Hood MI, Skaar EP. Nutritional immunity: transition metals at the pathogen-host interface. Nat Rev Microbiol. 2012 Jul 16;10(8):525-37.

[71] Andreini C, Bertini I, Cavallaro G, Holliday GL, Thornton JM. Metal ions in biological catalysis: from enzyme databases to general principles. J Biol Inorg Chem. 2008 Nov;13(8):1205-18.

[72] Sheldon JR, Laakso HA, Heinrichs DE. Iron Acquisition Strategies of Bacterial Pathogens. Microbiol Spectr. 2016 Apr;4(2)PubMed PMID: 27227297.

[73] Los FC, Randis TM, Aroian RV, Ratner AJ. Role of pore-forming toxins in bacterial infectious diseases. Microbiol Mol Biol Rev. 2013 Jun;77(2):173-207.

[74] Cescau S, Cwerman H, Létoffé S, Delepelaire P, Wandersman C, et al. Heme acquisition by hemophores. Biometals. 2007 Jun;20(3-4):603-13.

[75] Schalk IJ. Metal trafficking via siderophores in Gram-negative bacteria: specificities and characteristics of the pyoverdine pathway. J Inorg Biochem. 2008 May-Jun;102(5-6):1159-69.

[76] American Chemical Society International Historic Chemical Landmarks. Discovery and Development of Penicillin. `http://www.acs.org/content/acs/en/education/whatischemistry/landmarks/flemingpenicillin.html` (accessed May 31, 2016).

[77] Wright GD. The antibiotic resistome: the nexus of chemical and genetic diversity. Nat Rev Microbiol. 2007 Mar;5(3):175-86.

[78] Martinez JL. General principles of antibiotic resistance in bacteria. Drug Discov Today Technol. 2014 Mar;11:33-9.

[79] Martínez JL, Baquero F. Emergence and spread of antibiotic resistance: setting a parameter space. Ups J Med Sci. 2014 May;119(2):68-77.

[80] Levy SB, Marshall B. Antibacterial resistance worldwide: causes, challenges and responses. Nat Med. 2004 Dec;10(12 Suppl):S122-9.

[81] Finberg RW, Moellering RC, Tally FP, Craig WA, Pankey GA, et al. The importance of bactericidal drugs: future directions in infectious disease. Clin Infect Dis. 2004 Nov 1;39(9):1314-20.

[82] Alekshun MN, Levy SB. Commensals upon us. Biochem Pharmacol. 2006 Mar 30;71(7):893-900.

[83] Tavares A, Miragaia M, Rolo J, Coelho C, de Lencastre H. High prevalence of hospital-associated methicillin-resistant Staphylococcus aureus in the community in Portugal: evidence for the blurring of community-hospital boundaries. Eur J Clin Microbiol Infect Dis. 2013 Oct;32(10):1269-83.

[84] D'Costa VM, King CE, Kalan L, Morar M, Sung WW, et al. Antibiotic resistance is ancient. Nature. 2011 Aug 31;477(7365):457-61.

[85] Bhullar K, Waglechner N, Pawlowski A, Koteva K, Banks ED, et al. Antibiotic resistance is prevalent in an isolated cave microbiome. PLoS One. 2012;7(4):e34953.

[86] Wright GD, Poinar H. Antibiotic resistance is ancient: implications for drug discovery. Trends Microbiol. 2012 Apr;20(4):157-9.

[87] Benveniste R, Davies J. Aminoglycoside antibiotic-inactivating enzymes in actinomycetes similar to those present in clinical isolates of antibiotic-resistant bacteria. Proc Natl Acad Sci USA. 1973;70:2276–80.

[88] Baltz, R.H. Antibiotic discovery from actinomycetes: Will a renaissance follow the decline and fall? SIM News 55 (2005); 186–196

[89] D'Costa VM, McGrann KM, Hughes DW, Wright GD. Sampling the antibiotic resistome. Science. 2006;311:374–7.

[90] Forsberg KJ, Reyes A, Wang B, Selleck EM, Sommer MO, Dantas G. The shared antibiotic resistome of soil bacteria and human pathogens. Science. 2012;337:1107–11.

[91] Sommer MO, Dantas G, Church GM. Functional characterization of the antibiotic resistance reservoir in the human microflora. Science. 2009;325:1128–31.

[92] Levy, S. The Antibiotic Paradox: How Misuse of Antibiotics Destroys their Curative Powers (Perseus Cambridge, 2002).

[93] Davies J, Davies D. Origins and evolution of antibiotic resistance. Microbiol Mol Biol Rev. 2010 Sep;74(3):417-33.

[94] Wang X, Ryu D, Houtkooper RH, Auwerx J. Antibiotic use and abuse: a threat to mitochondria and chloroplasts with impact on research, health, and environment. Bioessays. 2015 Oct;37(10):1045-53.

[95] Van Boeckel TP, Brower C, Gilbert M, Grenfell BT, Levin SA, et al. Global trends in antimicrobial use in food animals. Proc Natl Acad Sci U S A. 2015 May 5;112(18):5649-54.

[96] Alekshun MN, Levy SB. Regulation of chromosomally mediated multiple antibiotic resistance: the mar regulon. Antimicrob Agents Chemother. 1997 Oct;41(10):2067-75.

[97] Gullberg E, Cao S, Berg OG, Ilbäck C, Sandegren L, et al. Selection of resistant bacteria at very low antibiotic concentrations. PLoS Pathog. 2011 Jul;7(7):e1002158.

[98] Gullberg E, Albrecht LM, Karlsson C, Sandegren L, Andersson DI. Selection of a multidrug resistance plasmid by sublethal levels of antibiotics and heavy metals. MBio. 2014 Oct 7;5(5):e01918-14.

[99] Baquero F, Coque TM. Widening the spaces of selection: evolution along sublethal antimicrobial gradients. MBio. 2014 Dec 9;5(6):e02270.

[100] Sandegren L. Selection of antibiotic resistance at very low antibiotic concentrations. Upsala Journal of Medical Sciences. 2014;119(2):103-107.

[101] Levy, S.B. Ecology of plasmids and unique DNA sequences. in Engineered Organisms in the Environment: Scientific Issues (eds. Halvorson, H.O., Pramer, D. & Rogul, M.) 180–190 (ASM Press, Washington DC, 1985).

[102] Berendonk TU, Manaia CM, Merlin C, Fatta-Kassinos D, Cytryn E, et al. Tackling antibiotic resistance: the environmental framework. Nat Rev Microbiol. 2015 May;13(5):310-7.

[103] Chang HH, Cohen T, Grad YH, Hanage WP, O'Brien TF, et al. Origin and proliferation of multiple-drug resistance in bacterial pathogens. Microbiol Mol Biol Rev. 2015 Mar;79(1):101-16.

[104] Blair JM, Webber MA, Baylay AJ, Ogbolu DO, Piddock LJ. Molecular mechanisms of antibiotic resistance. Nat Rev Microbiol. 2015 Jan;13(1):42-51.

[105] Gibson MK, Forsberg KJ, Dantas G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. ISME J. 2015 Jan;9(1):207-16.

[106] Abraham EP, Chain E. An enzyme from bacteria able to destroy penicillin 1940. Rev Infect Dis. 1988 Jul-Aug;10(4):677-8.

[107] Ambler RP. The structure of beta-lactamases. Philos Trans R Soc Lond B Biol Sci. 1980 May 16;289(1036):321-31.

[108] Kumar N, Radhakrishnan A, Wright CC, Chou TH, Lei HT, et al. Crystal structure of the transcriptional regulator Rv1219c of Mycobacterium tuberculosis. Protein Sci. 2014 Apr;23(4):423-32.

[109] Vetting MW, Hegde SS, Wang M, Jacoby GA, Hooper DC, et al. Structure of QnrB1, a plasmid-mediated fluoroquinolone resistance factor. J Biol Chem. 2011 Jul 15;286(28):25265-73.

[110] Depardieu F, Podglajen I, Leclercq R, Collatz E, Courvalin P. Modes and Modulations of Antibiotic Resistance Gene Expression. Clinical Microbiology Reviews. 2007;20(1):79-114.

[111] Arthur M, Molinas C, Courvalin P. The VanS-VanR two-component regulatory system controls synthesis of depsipeptide peptidoglycan precursors in Enterococcus faecium BM4147. J Bacteriol. 1992 Apr;174(8):2582-91.

[112] Marchand I, Damier-Piolle L, Courvalin P, Lambert T. Expression of the RND-type efflux pump AdeABC in Acinetobacter baumannii is regulated by the AdeRS two-component system. Antimicrob Agents Chemother. 2004 Sep;48(9):3298-304.

[113] Li X-Z, Nikaido H. Efflux-Mediated Drug Resistance in Bacteria: an Update. Drugs. 2009;69(12):1555-1623.

[114] McMurry L, Petrucci RE Jr, Levy SB. Active efflux of tetracycline encoded by four genetically different tetracycline resistance determinants in Escherichia coli. Proc Natl Acad Sci U S A. 1980 Jul;77(7):3974-7.

[115] Martinez JL, Sánchez MB, Martínez-Solano L, Hernandez A, Garmendia L, et al. Functional role of bacterial multidrug efflux pumps in microbial natural ecosystems. FEMS Microbiol Rev. 2009 Mar;33(2):430-49.

[116] Grkovic S, Brown MH, Skurray RA. Regulation of bacterial drug export systems. Microbiol Mol Biol Rev. 2002 Dec;66(4):671-701.

[117] Sun J, Deng Z, Yan A. Bacterial multidrug efflux pumps: mechanisms, physiology and pharmacological exploitations. Biochem Biophys Res Commun. 2014 Oct 17;453(2):254-67.

[118] Lubelski J, Konings WN, Driessen AJ. Distribution and physiology of ABC-type transporters contributing to multidrug resistance in bacteria. Microbiol Mol Biol Rev 2007; 71 (3): 463-76.

[119] Lu S, Zgurskaya HI. Role of ATP binding and hydrolysis in assembly of MacAB-TolC macrolide transporter. Molecular microbiology. 2012;86(5):1132-1143.

[120] Woebking B, Reuter G, Shilling RA, Velamakanni S, Shahi S, et al. Drug-lipid A interactions on the Escherichia coli ABC transporter MsbA. J Bacteriol. 2005 Sep;187(18):6363-9.

[121] Pao SS, Paulsen IT, Saier MH. Major Facilitator Superfamily. Microbiology and Molecular Biology Reviews. 1998;62(1):1-34.

[122] Ranaweera I, Shrestha U, Ranjana KC, et al. Structural comparison of bacterial multidrug efflux pumps of the major facilitator superfamily. Trends in cell & molecular biology. 2015;10:131-140.

[123] Kumar S, Mukherjee MM, Varela MF. Modulation of Bacterial Multidrug Resistance Efflux Pumps of the Major Facilitator Superfamily. International Journal of Bacteriology. 2013;2013:204141.

[124] Tanabe M, Szakonyi G, Brown KA, Henderson PJ, Nield J, et al. The multidrug resistance efflux complex, EmrAB from Escherichia coli forms a dimer in vitro. Biochem Biophys Res Commun. 2009 Mar 6;380(2):338-42.

[125] Andam CP, Fournier GP, Gogarten JP. Multilevel populations and the evolution of antibiotic resistance through horizontal gene transfer. FEMS Microbiol Rev. 2011 Sep;35(5):756-67.

[126] Soucy SM, Huang J, Gogarten JP. Horizontal gene transfer: building the web of life. Nat Rev Genet. 2015 Aug;16(8):472-82.

[127] Svara F, Rankin DJ. The evolution of plasmid-carried antibiotic resistance. BMC Evolutionary Biology. 2011;11:130.

[128] Nogueira T, Rankin DJ, Touchon M, Taddei F, Brown SP, Rocha EPC. Horizontal Gene Transfer of the Secretome Drives the Evolution of Bacterial Cooperation and Virulence. Current Biology. 2009;19(20):1683-1691.

[129] Trindade S, Sousa A, Xavier KB, Dionisio F, Ferreira MG, et al. Positive epistasis drives the acquisition of multidrug resistance. PLoS Genet. 2009 Jul;5(7):e1000578.

[130] Huddleston JR. Horizontal gene transfer in the human gastrointestinal tract: potential spread of antibiotic resistance genes. Infection and Drug Resistance. 2014;7:167-176.

[131] Trobos M, Lester CH, Olsen JE, Frimodt-Møller N, Hammerum AM. Natural transfer of sulphonamide and ampicillin resistance between Escherichia coli residing in the human intestine. J Antimicrob Chemother. 2009;63(1):80–86.

[132] Crémet L, Bourigault C, Lepelletier D, et al. Nosocomial outbreak of carbapenem-resistant Enterobacter cloacae highlighting the interspecies transferability of the blaOXA-48 gene in the gut flora. J Antimicrob Chemother. 2012;67(4):1041–1043.

[133] Karami N, Martner A, Enne VI, Swerkersson S, Adlerberth I, Wold AE. Transfer of an ampicillin resistance gene between two Escherichia coli strains in the bowel microbiota of an infant treated with antibiotics. J Antimicrob Chemother. 2007;60(5):1142–1145.

[134] Millard J, Ugarte-Gil C, Moore DA. Multidrug resistant tuberculosis. BMJ 2015;350:h882.

[135] Martínez JL, Baquero F. Interactions among strategies associated with bacterial infection: pathogenicity, epidemicity, and antibiotic resistance. Clin Microbiol Rev. 2002 Oct;15(4):647-79.

[136] Beceiro A, Tomás M, Bou G. Antimicrobial resistance and virulence: a successful or deleterious association in the bacterial world?. Clin Microbiol Rev. 2013 Apr;26(2):185-230.

[137] Soto SM. Role of efflux pumps in the antibiotic resistance of bacteria embedded in a biofilm. Virulence. 2013 Apr 1;4(3):223-9.

[138] Yeung AT, Bains M, Hancock RE. The sensor kinase CbrA is a global regulator that modulates metabolism, virulence, and antibiotic resistance in Pseudomonas aeruginosa. J Bacteriol. 2011 Feb;193(4):918-31.

[139] Baharoglu Z, Bikard D, Mazel D. Conjugative DNA transfer induces the bacterial SOS response and promotes antibiotic resistance development through integron activation. PLoS Genet. 2010 Oct 21;6(10):e1001165.

[140] Davies J, Spiegelman GB, Yim G. The world of subinhibitory antibiotic concentrations. Curr Opin Microbiol. 2006 Oct;9(5):445-53.

[141] Hall RM, Collis CM. Mobile gene cassettes and integrons: capture and spread of genes by site-specific recombination. Mol Microbiol. 1995 Feb;15(4):593-600.

[142] Gillings MR. Integrons: Past, Present, and Future. Microbiology and Molecular Biology Reviews: MMBR. 2014;78(2):257-277.

[143] Guerra B, Soto S, Helmuth R, Mendoza MC. Characterization of a self-transferable plasmid from Salmonella enterica serotype typhimurium clinical isolates carrying two integron-borne gene cassettes together with virulence and drug resistance genes. Antimicrob Agents Chemother. 2002 Sep;46(9):2977-81.

[144] Mazel D, Dychinco B, Webb VA, Davies J. A distinctive class of integron in the Vibrio cholerae genome. Science. 1998 Apr 24;280(5363):605-8.

[145] Ghigo JM. Natural conjugative plasmids induce bacterial biofilm development. Nature. 2001 Jul 26;412(6845):442-5.

[146] Molin S, Tolker-Nielsen T. Gene transfer occurs with enhanced efficiency in biofilms and induces enhanced stabilisation of the biofilm structure. Curr Opin Biotechnol. 2003 Jun;14(3):255-61.

[147] Hayes CS, Aoki SK, Low DA. Bacterial contact-dependent delivery systems. Annu Rev Genet. 2010;44:71-90.

[148] Parsek MR, Singh PK. Bacterial biofilms: an emerging link to disease pathogenesis. Annu Rev Microbiol. 2003;57:677-701.

[149] De Kievit TR, Parkins MD, Gillis RJ, Srikumar R, Ceri H, Poole K, et al. Multidrug efflux pumps: expression patterns and contribution to antibiotic resistance in Pseudomonas aeruginosa biofilms. Antimicrob Agents Chemother 2001; 45:1761-70.

[150] Ito A, Taniuchi A, May T, Kawata K, Okabe S. Increased antibiotic resistance of Escherichia coli in mature biofilms. Appl Environ Microbiol 2009; 75:4093-100.

[151] Delmont TO, Malandain C, Prestat E, Larose C, Monier JM, et al. Metagenomic mining for microbiologists. ISME J. 2011 Dec;5(12):1837-43.

[152] Hugenholtz P, Goebel BM, Pace NR. Impact of Culture-Independent Studies on the Emerging Phylogenetic View of Bacterial Diversity. Journal of Bacteriology. 1998;180(18):4765-4774.

[153] Amann RI, Ludwig W, Schleifer KH. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiological Reviews. 1995;59(1):143-169.

[154] Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR, Stahl DA. Microbial ecology and evolution: a ribosomal RNA approach. Annu Rev Microbiol. 1986;40:337-65.

[155] Eisen JA. Environmental Shotgun Sequencing: Its Potential and Challenges for Studying the Hidden World of Microbes. PLoS Biology. 2007;5(3):e82.

[156] Freilich S, Zarecki R, Eilam O, Segal ES, Henry CS, et al. Competitive and cooperative metabolic interactions in bacterial communities. Nat Commun. 2011 Dec 13;2:589.

[157] Ponomarova O, Patil KR. Metabolic interactions in microbial communities: untangling the Gordian knot. Curr Opin Microbiol. 2015 Oct;27:37-44.

[158] Rosenthal AZ, Matson EG, Eldar A, Leadbetter JR. RNA-seq reveals cooperative metabolic interactions between two termite-gut spirochete species in co-culture. ISME J. 2011 Jul;5(7):1133-42.

[159] Rankin DJ, Ginty SEM, Nogueira T, et al. Bacterial cooperation controlled by mobile elements: kin selection and infectivity are part of the same process. Heredity. 2011;107(3):279-281.

[160] Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, et al. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. 2008 Sep 19;9:386.

[161] Markowitz VM, Chen I-MA, Palaniappan K, et al. IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Research. 2012;40(Database issue):D115-D122.

[162] Mitchell A, Bucchini F, Cochrane G, et al. EBI metagenomics in 2016 - an expanding and evolving resource for the analysis and archiving of metagenomic data. Nucleic Acids Research. 2016;44(Database issue):D595-D603.

[163] Wilke A, Bischof J, Gerlach W, Glass E, Harrison T, et al. The MG-RAST metagenomics database and portal in 2015. Nucleic Acids Res. 2016 Jan 4;44(D1):D590-4.

[164] Keegan KP, Glass EM, Meyer F. MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. Methods Mol Biol. 2016;1399:207-33.

[165] Wilke A, Bischof J, Harrison T, Brettin T, D'Souza M, et al. A RESTful API for accessing microbial community data for MG-RAST. PLoS Comput Biol. 2015 Jan;11(1):e1004008.

[166] Mulcahy-O'Grady H, Workentine ML. The Challenge and Potential of Metagenomics in the Clinic. Front Immunol. 2016;7:29.

[167] Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, et al. Human gut microbiome viewed across age and geography. Nature. 2012 May 9;486(7402):222-7.

[168] Koenig JE, Spor A, Scalfone N, et al. Succession of microbial consortia in the developing infant gut microbiome. Proceedings of the National Academy of Sciences of the United States of America. 2011;108(Suppl 1):4578-4585.

[169] Barroso-Batista J, Demengeot J, Gordo I. Adaptive immunity increases the pace and predictability of evolutionary change in commensal gut bacteria. Nature Communications. 2015;6:8945.

[170] Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. Nat Rev Genet. 2012 Mar 13;13(4):260-70.

[171] Moore AM, Patel S, Forsberg KJ, et al. Pediatric Fecal Microbiota Harbor Diverse and Novel Antibiotic Resistance Genes. Tse H, ed. PLoS ONE. 2013;8(11):e78822.

[172] Clemente JC, Pehrsson EC, Blaser MJ, et al. The microbiome of uncontacted Amerindians. Scienc Advances. 2015;1(3):e1500183.

[173] Moore AM, Ahmadi S, Patel S, et al. Gut resistome development in healthy twin pairs in the first year of life. Microbiome. 2015;3:27.

[174] Pehrsson EC, Tsukayama P, Patel S, Mejía-Bautista M, Sosa-Soto G, et al. Interconnected microbiomes and resistomes in low-income human habitats. Nature. 2016 May 11;533(7602):212-6.

[175] Fang H, Wang H, Cai L, Yu Y. Prevalence of antibiotic resistance genes and bacterial pathogens in long-term manured greenhouse soils as revealed by metagenomic survey. Environ Sci Technol. 2015 Jan 20;49(2):1095-104.

[176] Li B, Ju F, Cai L, Zhang T. Profile and Fate of Bacterial Pathogens in Sewage Treatment Plants Revealed by High-Throughput Metagenomic Approach. Environ Sci Technol. 2015 Sep 1;49(17):10492-502.

[177] Ju F, Li B, Ma L, Wang Y, Huang D, et al. Antibiotic resistance genes and human bacterial pathogens: Co-occurrence, removal, and enrichment in municipal sewage sludge digesters. Water Res. 2016 Mar 15;91:1-10.

[178] Pearson WR. An introduction to sequence similarity ("homology") searching. Curr Protoc Bioinformatics. 2013 Jun;Chapter 3:Unit3.1.

[179] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970 Mar;48(3):443-53.

[180] Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981 Mar 25;147(1):195-7.

[181] Sellers PH. On the theory and computation of evolutionary distances. SIAM.J. Appl. Math. 1974 Jun;26(4):787-93.

[182] Gotoh O. An improved algorithm for matching biological sequences. J Mol Biol. 1982 Dec 15;162(3):705-8.

[183] Altschul SF, Erickson BW. Optimal sequence alignment using affine gap costs. Bull Math Biol. 1986;48(5-6):603-16.

[184] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10.

[185] Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proceedings of the National Academy of Sciences of the United States of America. 1990;87(6):2264-2268.

[186] Karlin S, Altschul SF. Applications and statistics for multiple high-scoring segments in molecular sequences. Proceedings of the National Academy of Sciences of the United States of America. 1993;90(12):5873-5877.

[187] Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 1997;25(17):3389-3402.

[188] Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A. 1988 Apr;85(8):2444-8.

[189] Pearson WR. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. Genomics. 1991 Nov;11(3):635-50.

[190] Kent WJ. BLAT—The BLAST-Like Alignment Tool. Genome Research. 2002;12(4):656-664.

[191] Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010 Oct 1;26(19):2460-1.

[192] Johnson LS, Eddy SR, Portugaly E. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics. 2010 Aug 18;11:431.

[193] Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015 Jan;12(1):59-60.

[194] Watson JD, Baker TA, Bell SP, Gann A, Levine M, Oosick R. Molecular Biology of the Gene. Chapter 15. San Francisco: Pearson/Benjamin Cummings (2008).

[195] Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. Nature reviews Genetics. 2011;12(1):32-42.

[196] Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Research. 2010;38(20):e191.

[197] Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Weizhong Li & Adam Godzik. Bioinformatics, (2006) 22:1658-9.

[198] Chen L, Xiong Z, Sun L, Yang J, Jin Q. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. Nucleic Acids Res. 2012 Jan;40(Database issue):D641-5.

[199] Liu B, Pop M. ARDB–Antibiotic Resistance Genes Database. Nucleic Acids Res. 2009 Jan;37(Database issue):D443-7.

[200] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;57(1):289–300.

[201] Welch BL. The generalization of student's problems when several different population variances are involved. Biometrika. 1947;34(1-2):28-35.

[202] Ruxton GD. The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. Behavioral Ecology. 2006 Jul;17(4):688–690.

[203] Fagerland MW. t-tests, non-parametric tests, and large studies—a paradox of statistical practice? BMC Medical Research Methodology. 2012;12:78.

[204] Martínez JL, Coque TM, Baquero F. What is a resistance gene? Ranking risk in resistomes. Nat Rev Microbiol. 2015 Feb;13(2):116-23.

[205] Martínez JL. Antibiotics and antibiotic resistance genes in natural environments. Science. 2008 Jul 18;321(5887):365-7.

[206] Fitzpatrick D, Walsh F. Antibiotic resistance genes across a wide variety of metagenomes. FEMS Microbiol Ecol. 2016 Feb;92(2)PubMed PMID: 26738556.

[207] Ghosh TS, Gupta SS, Nair GB, Mande SS. In silico analysis of antibiotic resistance genes in the gut microflora of individuals from diverse geographies and age-groups. PLoS One. 2013;8(12):e83823.

[208] Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. Diversity, stability and resilience of the human gut microbiota. Nature. 2012 Sep 13;489(7415):220-30.

[209] Ghosh TS, Gupta SS, Bhattacharya T, Yadav D, Barik A, et al. Gut microbiomes of Indian children of varying nutritional status. PLoS One. 2014;9(4):e95547.

[210] Subramanian S, Huq S, Yatsunenko T, et al. Persistent Gut Microbiota Immaturity in Malnourished Bangladeshi Children. Nature. 2014;510(7505):417-421.

[211] Alcoba G, Kerac M, Breysse S, Salpeteur C, Galetto-Lacour A, et al. Do children with uncomplicated severe acute malnutrition need antibiotics? A systematic review and meta-analysis. PLoS One. 2013;8(1):e53184.

[212] Makoka MH, Miller WC, Hoffman IF, Cholera R, Gilligan PH, et al. Bacterial infections in Lilongwe, Malawi: aetiology and antibiotic resistance. BMC Infect Dis. 2012 Mar 21;12:67.

[213] Nogueira T, Touchon M, Rocha EP. Rapid evolution of the sequences and gene repertoires of secreted proteins in bacteria. PLoS One. 2012;7(11):e49403.

[214] Bisognano C, Kelley WL, Estoppey T, Francois P, Schrenzel J, et al. A recA-LexA-dependent pathway mediates ciprofloxacin-induced fibronectin binding in Staphylococcus aureus. J Biol Chem. 2004 Mar 5;279(10):9064-71.

[215] Hood MI, Jacobs AC, Sayood K, Dunman PM, Skaar EP. Acinetobacter baumannii increases tolerance to antibiotics in response to monovalent cations. Antimicrob Agents Chemother. 2010 Mar;54(3):1029-41.

[216] Drevinek P, Holden MT, Ge Z, Jones AM, Ketchell I, et al. Gene expression changes linked to antimicrobial resistance, oxidative stress, iron depletion and retained motility are observed when Burkholderia cenocepacia grows in cystic fibrosis sputum. BMC Infect Dis. 2008 Sep 19;8:121.